Stochastic process customer lifetime value models with time-varying covariates

David M. Harman

University of Iowa

Follow this and additional works at: https://ir.uiowa.edu/etd

Part of the Business Administration, Management, and Operations Commons

Copyright © 2016 David M. Harman

This dissertation is available at Iowa Research Online: https://ir.uiowa.edu/etd/2221

Recommended Citation

https://doi.org/10.17077/etd.zjhm23na

Follow this and additional works at: https://ir.uiowa.edu/etd

Part of the Business Administration, Management, and Operations Commons
STOCHASTIC PROCESS
CUSTOMER LIFETIME VALUE MODELS
WITH TIME-VARYING COVARIATES

by

David M. Harman

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Business Administration in the Graduate College of The University of Iowa

December 2016

Thesis Supervisor: Professor Thomas S. Gruca
This is to certify that the Ph.D. thesis of

David M. Harman

has been approved by the Examining Committee for
the thesis requirement for the Doctor of Philosophy degree
in Business Administration at the December 2016 graduation.

Thesis Committee:

Thomas S. Gruca, Thesis Supervisor

Gary J. Russell

Sanghak Lee

William Hedgock

Mary Kathryn Cowles
"Essentially, all models are wrong, but some are useful."

George E. P. Box, with Norman R. Draper
ABSTRACT

Customer lifetime value (CLV) is a forecasted expectation of how much value a customer is likely to contribute to the firm in the future either before they leave the firm or within a specified window of time. There are two customer behavioral components of CLV that represent a particular modeling challenge: 1) how many transactions can we expect from a customer in the future, and 2) how likely is it that the customer remains active. A major portion of the academic marketing literature for CLV models is devoted to stochastic process CLV models, e.g. the Pareto/NBD, that combine separate transaction and attrition processes within a model in order to provide estimates for these two behavioral components of CLV. Stochastic process CLV models are valuable managerial tools because they are able to provide forward-looking estimates of transaction patterns and customer churn when the event of a customer leaving is unobservable, which is typical for most noncontractual consumer goods and services.

The CLV modeling literature has for the most part maintained its original assumption that the number of customer transactions follows a stable process. A stable process implies that over time periods of equal length, say, the last four weeks and the next four weeks, a customer is expected to transact roughly the same amount of times in each time period. But there are many categories of noncontractual goods and services where a stable transaction rate assumption is violated, particularly categories with seasonal transaction patterns. CLV model estimates are further biased when there is an excess of customers with no repeat transactions.

To address these modeling challenges, I develop a generalized CLV modeling framework that combines three elements necessary to reduce bias in model estimates: 1) the incorporation of time-varying covariates to model nonstationary, particularly seasonal, transaction data, 2) a zero-inflated model specification for customers with no repeat transactions, and 3) generalizes to different transaction process distributions to better fit diverse customer transaction patterns. The CLV modeling framework developed in this thesis will provide marketing professionals with a
better approach to estimate the value of their customer base, a critical CRM application for businesses that do not have a direct way to measure when a customer leaves (e.g., cancelling a contract).
PUBLIC ABSTRACT

Customer lifetime value (CLV) is a forecasted expectation of the future value of a customer to the firm. There are two customer behavioral components of CLV that represent a particular modeling challenge: 1) how many transactions we expect from a customer in the future, and 2) how likely it is the customer remains active. Existing CLV models like the Pareto/NBD are valuable managerial tools because they are able to provide forward-looking estimates of transaction patterns and customer churn when the event of a customer leaving is unobservable, which is typical for most noncontractual goods and services.

The CLV model literature has for the most part maintained its original assumption that the number of customer transactions follows a stable transaction process. Yet there are many categories of noncontractual goods and services where the stable transaction rate assumption is violated, particularly seasonal purchase patterns. CLV model estimates are further biased when there is an excess of customers with no repeat transactions.

To address these modeling challenges, within this thesis I develop a generalized CLV modeling framework that combines three elements necessary to reduce bias in model estimates: 1) the incorporation of time-varying covariates to model data with transaction rates that change over time, 2) a zero-inflated model specification for customers with no repeat transactions, and 3) generalizes to different transaction process distributions to better fit diverse customer transaction patterns. This CLV modeling framework provides firms better estimates of the future activity of their customers, a critical CRM application.
TABLE OF CONTENTS

List of Figures ........................................................................................................................................ ix
List of Tables ........................................................................................................................................... xii
Chapter 1: Introduction ........................................................................................................................ 1
Chapter 2: Stochastic Process Models in Marketing .............................................................................. 6
  An Historical Defense of Stochastic Process Models ................................................................. 9
  Transaction Timing Models ............................................................................................................. 12
  Hazard Function Transaction Models ......................................................................................... 19
  Stochastic Brand Choice Models ................................................................................................. 23
  Non-Transaction Stochastic Process Marketing Models ........................................................... 29
Chapter 3: Stochastic Process CLV Models ......................................................................................... 32
  The Generalized Structure of Stationary Stochastic Process CLV Models .............................. 34
  Why Time-Varying Covariates Are Not Common in the CLV Literature ................................. 39
  The Uncorrected Individual Likelihood Function Leads to Biased Parameter and
  Model Estimates .......................................................................................................................... 41
  Seasonality Produces Biased CLV Model Estimates ................................................................. 41
  Zero-Repurchasing Customers Bias Parameter Estimates ......................................................... 43
  Which Transaction Distribution Is “Best?” .................................................................................... 44
  CLV Model Likelihood Discussion ............................................................................................... 45
    Stochastic Process CLV Model Literature Review ................................................................. 45
  Applied and Slightly Modified Stochastic Process CLV Models ............................................... 48
  New Stochastic Process CLV Models After the Pareto/NBD ................................................. 51
  More Complex Stochastic Process CLV Models ......................................................................... 57
  CLV Models with Time-Varying Covariates and Endogeneity Correction ............................... 60
Chapter 4: Different Model Assumptions and Dataset Description .................................................. 65
LIST OF FIGURES

3.1: Example of Observed Purchase Data .................................................................................. 35
4.1: Soup Cohort Transactions by Week .................................................................................... 71
4.2: Number of Soup Customers by Number of Repeat Transactions, 4 Years Observed .......... 72
4.3: Average Intertransaction Times by Soup Customers with One or More Repeat Transactions .......................................................................................................................... 72
5.1: Observed and Holdout Data ................................................................................................. 86
5.2: Simulated Transactions Example, Erlang-2 Transaction Process, $\gamma = 3$ .................... 89
5.3: Proportion of Zero-Repeat Customers .................................................................................. 92
5.4a: Recovery of Zero-Repeat Customers, Exponential Transactions (Bias) ......................... 93
5.4b: Recovery of Zero-Repeat Customers, Exponential Transactions (RMSE) ...................... 94
5.5a: Recovery of Zero-Repeat Customers, Erlang-2 Transactions (Bias) ............................... 95
5.5b: Recovery of Zero-Repeat Customers, Erlang-2 Transactions (RMSE) ......................... 96
5.6a: Recovery of Zero-Repeat Customers, Erlang-3 Transactions (Bias) ............................... 97
5.6b: Recovery of Zero-Repeat Customers, Erlang-3 Transactions (RMSE) ......................... 98
5.7a: Recovery of $E[p]$, Exponential Transactions (Bias) ........................................................ 101
5.7b: Recovery of $E[p]$, Exponential Transactions (RMSE) ..................................................... 102
5.8a: Recovery of $E[p]$, Erlang-2 Transactions (Bias) .............................................................. 103
5.8b: Recovery of $E[p]$, Erlang-2 Transactions (RMSE) .......................................................... 104
5.9a: Recovery of $E[p]$, Erlang-3 Transactions (Bias) .............................................................. 105
5.9b: Recovery of $E[p]$, Erlang-3 Transactions (RMSE) .......................................................... 106
5.10: Mean Proportion of Customers Still Active ........................................................................ 108
5.11a: Recovery of $E[P(Alive)]$, Exponential Transactions (Bias) ............................................ 109
5.11b: Recovery of $E[P(Alive)]$, Exponential Transactions (RMSE) ........................................ 110
5.12a: Recovery of $E[P(Alive)]$, Erlang-2 Transactions (Bias) ............................................... 111
5.12b: Recovery of E[P(Alive)], Erlang-2 Transactions (RMSE) .......................... 112
5.13a: Recovery of E[P(Alive)], Erlang-3 Transactions .................................. 113
5.13b: Recovery of E[P(Alive)], Erlang-3 Transactions (RMSE) ...................... 114
5.14: Recovery of γ, Exponential, Erlang-2, and Erlang-3 Transactions .......... 117
5.15a: Recovery of E[λ], Exponential Transactions (Bias) ............................ 121
5.15b: Recovery of E[λ], Exponential Transactions (RMSE) .......................... 122
5.16a: Recovery of E[λ], Erlang-2 Transactions (Bias) ................................. 123
5.16b: Recovery of E[λ], Erlang-2 Transactions (RMSE) .............................. 124
5.17a: Recovery of E[λ], Erlang-3 Transactions (Bias) ................................. 125
5.17b: Recovery of E[λ], Erlang-3 Transactions (RMSE) .............................. 126
5.18: Mean Future Transactions ..................................................................... 129
5.19a: Recovery of Future Transactions, Exponential Transactions (Bias) ....... 130
5.19b: Recovery of Future Transactions, Exponential Transactions (RMSE) .... 131
5.20a: Recovery of Future Transactions, Erlang-2 Transactions (Bias) .......... 132
5.20b: Recovery of Future Transactions, Erlang-2 Transactions (RMSE) ....... 133
5.21a: Recovery of Future Transactions, Erlang-3 Transactions (Bias) .......... 134
5.21b: Recovery of Future Transactions, Erlang-3 Transactions (RMSE) ....... 135
5.22: Newton-Raftery Estimates, Exponential, Erlang-2, and Erlang-3 Transactions .......................... 138
5.23: Estimated Proportion of Zero-Repeat Soup Customers at Model Fitting .... 143
5.24: Estimated Soup E[P(Alive)] at Model Fitting ...................................... 145
5.25: Forecasted Soup Transactions, 104 Weeks Forward Looking ......... 146
5.26: Estimated Cosine Trend Seasonal Correction Parameters .................. 149
6.1: Simulated Transactions Example, Exponential Transaction Process .... 152
6.2: Weibull Shape Parameter Example MCMC Chains ............................ 154
6.3: Model Newton-Raftery Score Difference in Proportions ..................... 167
6.4: Proportional Size of Class One, Two-Class Model ................................................................. 168
6.5: Covariate Parameter Values, Two-Class Model ................................................................. 170
6.6: Estimated P(Alive) in Proportion to Actual, by Model ...................................................... 172
6.7: Estimated Future Transactions in Proportion to Actual, by Model ...................................... 173
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>CLV Articles with Time-Varying Covariates</td>
<td>61</td>
</tr>
<tr>
<td>5.1</td>
<td>Recovery of Zero-Repeat Customers, RMSE</td>
<td>99</td>
</tr>
<tr>
<td>5.2</td>
<td>Recovery of $E[p]$, RMSE</td>
<td>107</td>
</tr>
<tr>
<td>5.3</td>
<td>Recovery of $E[P(\text{Alive})]$, RMSE</td>
<td>115</td>
</tr>
<tr>
<td>5.4</td>
<td>Recovery of $E[\lambda]$, RMSE</td>
<td>127</td>
</tr>
<tr>
<td>5.5</td>
<td>Recovery of Future Transactions, RMSE</td>
<td>136</td>
</tr>
<tr>
<td>5.6</td>
<td>Estimated Soup $E[P(\text{Alive})]$, RMSE and MAPE</td>
<td>145</td>
</tr>
<tr>
<td>5.7</td>
<td>Forecasted Soup Transactions, RMSE and MAPE</td>
<td>147</td>
</tr>
<tr>
<td>6.1</td>
<td>Average Bivariate Normal Model Estimates of Correlation Between Simulated Transaction and Attrition Rate Parameters by Prior</td>
<td>157</td>
</tr>
<tr>
<td>6.2</td>
<td>RMSE of Average Exponential and Bivariate Normal Model Estimates</td>
<td>158</td>
</tr>
<tr>
<td>6.3</td>
<td>MAPE of Average Exponential and Bivariate Normal Model Estimates</td>
<td>158</td>
</tr>
<tr>
<td>C.1</td>
<td>Convergence Tests Passed for MCMC Draws of $\omega$</td>
<td>191</td>
</tr>
<tr>
<td>C.2</td>
<td>Convergence Tests Passed for MCMC Draws of $E[p]$</td>
<td>192</td>
</tr>
<tr>
<td>C.3</td>
<td>Convergence Tests Passed for MCMC Draws of $E[P(\text{Alive})]$</td>
<td>193</td>
</tr>
<tr>
<td>C.4</td>
<td>Convergence Tests Passed for MCMC Draws of $\gamma$</td>
<td>193</td>
</tr>
<tr>
<td>C.5</td>
<td>Convergence Tests Passed for MCMC Draws of $E[\lambda]$</td>
<td>194</td>
</tr>
<tr>
<td>C.6</td>
<td>Convergence Tests Passed for MCMC Draws of $E[\text{Future Transactions}]$</td>
<td>195</td>
</tr>
</tbody>
</table>
CHAPTER 1: INTRODUCTION

Customer lifetime value (CLV) is the expectation of the future value of an individual customer to the firm. This is an important metric for managers at multiple levels of the organization. At the strategic level, the aggregated CLV of a firm’s customer base – commonly termed the firm’s customer equity – is a key driver of the firm’s value to shareholders (Gupta, Lehmann, and Stuart 2004; Gupta et al. 2006; Kumar and Shah 2009). As part of the managerial decision making process, estimates of customer lifetime values are used for resource allocation decisions including service quality, advertising, direct marketing, and distribution (Bolton, Lemon, and Verhoef 2004; Kumar et al. 2008). Because of the linkage between CLV and firm value, an accurate estimate of CLV is needed so managers can assess the impact of different marketing programs. Systematic over or underestimates of CLV may lead managers to make poor judgments regarding the effectiveness of various marketing programs. For managers to make decisions that will improve customer lifetime value, analysts need to provide managers with accurate estimates of what customers will do in the future.

There are two customer behavioral components of CLV that are a particular modeling challenge: 1) how many transactions can we expect from a customer in the future, and 2) how likely is the customer to remain active. When we are able to observe when a customer leaves, usually in a contractual context such as the cancellation of wireless phone service or a magazine subscription, we can calibrate a model on observed cancellation data to estimate how long existing customers will remain active (e.g. Bolton 1998; Borle, Singh, and Jain 2008). However, with most noncontractual consumer goods and services, the event of a customer leaving is unobservable in the data, and the firm does not know when their customers choose to leave. In these noncontractual settings, CLV models like the Pareto/NBD (Schmittlein, Morrison, and Colombo 1987) and the BG/NBD (Fader, Hardie, and Lee 2005a) provide estimates for the number of future transactions and the probability the customer remains active even though the act
of a customer leaving is unobservable. This is accomplished by combining a stochastic process that models the number of customer transactions along with a second, latent stochastic process to represent whether or not a customer has left.

In the current literature, e.g. Schweidel and Knox (2013), stochastic process CLV models are often referred to as latent attrition models in acknowledgment of this latent attrition stochastic process. Latent attrition models focus on modeling the transaction and attrition processes, and, more often than not, choose to ignore the monetary component of CLV. This is in recognition that customer transactions and attrition are far more difficult to forecast than what a customer will spend, which over multiple future transactions is very close to the average of their past spending per transaction. This thesis focuses specifically on latent attrition models, and does not include the monetary component of CLV in the models herein. However, this thesis does maintain the common vernacular term of CLV model that a wider set of readers are more likely to recognize.

For the most part, the stochastic process CLV model literature has maintained the original assumption of the Pareto/NBD that customer transactions follow a stationary stochastic process to model the future expected number of transactions from a customer base. With the stationary transaction assumption, the models assume that the observed amounts of time that pass between each transaction for a customer are represented by the same joint distribution for that customer so that, over multiple transactions, the time between transactions is random with respect to that joint distribution for that customer. But there are many categories of noncontractual goods and services where the stationary transaction assumption is violated, particularly categories with a seasonal transaction pattern. This includes holiday gifts, warm weather gardening supplies, condiments ahead of summer “barbeque” holidays, non-profit donations, etc. If the transaction data being modeled is nonstationary, then current stationary models will produce biased estimates of future customer activity that vary based on when within a seasonal cycle the model is estimated. This bias was anticipated with existing stationary CLV models, as even Schmittlein, Morrison, and Colombo (1987, pg. 3) recognize, "The [stationary] Poisson assumption would not
be as good with catalog purchases since the new editions of the catalog may be mailed out on a very regular schedule."

The challenge posed by this problem is that CLV models are difficult to parameterize on a set of explanatory covariates that could correct for nonstationary seasonal cycles. Furthermore, we as researchers often look to the fitted parameters on covariates in a model to provide a deterministic explanation of why we are observing different patterns in the data. Stochastic process models are fitted to data observed over a period of time which can be relatively straightforward when the data are stationary. But if the data are nonstationary and consumer activity that varies with time is inherently nonstationary, it is far more difficult to incorporate and parameterize time-varying covariates in stochastic process models compared to, say, logit choice models. The need to add covariates to stochastic process CLV models was recognized by Schmittlein, Morrison, and Colombo (1987), and they provided a general explanation of how to add time-invariant covariates to their Pareto/NBD. Fader and Hardie (2007) describe more specifically how to add time-invariant covariates to the Pareto/NBD model and their BG/NBD model. However, a generalizable method of adding time-varying covariates is needed for stochastic process CLV models as observed customer behaviors and marketing efforts targeted at customers are rarely invariant over time.

In this thesis, I incorporate time-varying covariates into the transaction process through a proportional hazards formulation. Proportional hazards models have been commonly used in the marketing literature to model event timing (e.g., Jain and Vilcassim 1991, Helsen and Schmittlein 1993, Fader and Hardie 2007). Schweidel and Knox (2013) were the first to incorporate time-varying covariates into a Bernoulli transaction process within a CLV model. However, the methodology I develop in this thesis generalizes to any transaction distribution with a hazard function that is a function of time.

This generalization is important. Platzer (2008) and Platzer and Reutterer (2016) demonstrate the specification of the transaction distribution can improve CLV model estimates.
Since the marketing event timing literature has demonstrated that the transaction distribution can vary by product category, CLV models must be able to accommodate different transaction distributions.

My CLV modeling framework also incorporates a zero-inflation model specification. Zero-inflated models are a well-known modeling method for count data with an excess of zeros. They also have a long history in the modeling literature, e.g., Morrison (1969). The zero-inflated model specification reduces parameter bias and improves estimates of future customer behavior to the point where we get more accurate estimates of whether the customer is still active. The need for a zero-inflated model arises out of the individual likelihood function (see Chapter 3). This addition to the modeling framework addresses a criticism by Wübben and von Wangenheim (2008) that unsophisticated managerial heuristics are at least as effective as stochastic process CLV models at identifying inactive customers. To my knowledge, my CLV modeling framework is the first use of a zero-inflated specification within a lost-for-good CLV model.

The purpose of this thesis is to directly address the gap in the stochastic process CLV model literature that currently does not have a general framework for adding time-varying covariates to account for nonstationary customer transaction patterns. This thesis will develop a new CLV modeling framework that generalizes how to incorporate time-varying covariates that will lead to less biased estimates of CLV compared to existing stochastic process CLV models in the presence of nonstationary patterns in the data, e.g., seasonal purchasing.

There is a large body of literature that I will be reviewing in this thesis that underlies existing CLV models, a fair amount of it largely forgotten, i.e. rarely referenced in current marketing quantitative research. As a result, I am eschewing the three-essay thesis format in favor of a chapter format that underlies a cohesive thesis in order to draw on that commonality throughout. The rest of this thesis is organized as follows. Chapter 2 provides an historical review of stochastic process and event timing models in the marketing literature. Fader, Hardie, and Sen (2014) cover much of the history of early stochastic process models, but they give little
weight to some of the papers I find significant to build upon. There is new ground to be covered from reviewing the stochastic choice models of this time period, at a minimum to provide more nuance to how this branch of the marketing literature evolved with Guadagni and Little’s (1983) pivotal paper after which marketing researchers focused more on logit choice modeling. Chapter 3 reviews the stochastic process CLV modeling literature. This chapter includes a discussion of the generalization of the likelihood function of an individual customer in a CLV model. Chapter 4 summarizes the different CLV models I will be using for comparative purposes in the modeling applications and a description of the soup transaction dataset that will be utilized in the empirical application that follows. Chapters 5 is where I derive my CLV modeling framework that includes time-varying covariates and a zero-inflated model specification, and generalizes the transaction distribution. The CLV modeling framework is demonstrated with simulated data as well as soup transaction data. Chapter 6 extends my CLV modeling framework with latent classes. This extension does have limitations with the parameter recovery of simulated data. The chapter includes discussion of problems with the latent class model as well as documents other boundary conditions where there are known inference issues with my CLV modeling framework. Chapter 7 concludes with a discussion of what this thesis adds to our knowledge.

My CLV modeling framework pushes the boundary of stochastic process CLV models to the point where these models can provide reasonably close estimates of future customer behaviors in terms of whether they remain active and how many transactions can we expect from them, making these models more useful for managers. This is the primary contribution of this work. But in the course of conducting the research underlying this thesis, there remain quite a few unanswered questions that will be discussed further in Chapters 5 and 6. It is my hope these remaining questions will be the start of an active research program, not the end of one.
CHAPTER 2: STOCHASTIC PROCESS MODELS IN MARKETING

Prior to Guadagni and Little's (1983) demonstration of the usefulness of the logit model to marketing research, stochastic process models represented a major branch of marketing science research including models of purchase timing as well as brand choice (Massy, Montgomery, and Morrison 1970). Prominent examples include the negative binomial distribution (hereafter NBD), first introduced to the marketing literature by Ehrenberg (1959) and the Dirichlet (multivariate Beta) distribution introduced to the marketing literature by Chatfield and Goodhardt (1975). The NBD has been used to model purchase rates, market penetration, and more. The Beta-Bernoulli distribution was a more common choice model prior to the mid-1970's, but the Dirichlet allowed researchers to model multivariate brand choice, and early extensions included modeling the dependence between brand choice and purchase frequency (Shoemaker, Staelin, Kadane, and Shoaf 1977). During the mid-1980's, more complex stochastic models were being proposed to simultaneously model both brand choice and purchase timing, for example, as when Goodhardt, Ehrenberg, and Chatfield (1984) combined a Dirichlet-multinomial model of brand choice with the NBD to model purchase timing.

Stochastic process models are useful at describing what is going on in the marketplace, less so at explaining why. The challenge behind these models is they are far more difficult to parameterize on a set of explanatory covariates that could provide a deterministic explanation of why we are observing different patterns in the data. Difficult, but still possible. Both Wagner and Taudes (1986, 1991) and Gupta (1991) have demonstrated different stochastic process marketing models that can be parameterized on time-varying covariates. Later research in the 1990’s further developed the parameterization of stochastic process models with time-varying marketing covariates, either through parameterizing the parameters, i.e. as a hierarchical model, or through the hazard function of the stochastic process.
Unfortunately, these efforts to include explanatory covariates in stochastic process models took place after Guadagni and Little's (1983) pivotal paper, which helps to explain why these models have faded from use in the marketing literature since the 1980's (Fader, Hardie, and Sen 2014). Though not mentioned in the literature, another likely reason for why logit models displaced stochastic process models is the logit model's representation of a consumer's utility as a latent variable, thus tying logit choice models directly to economic theory. While there are instances of similar latent representations of utility in stochastic process models of brand choice, e.g. Bass, Jeuland, and Wright (1976), these are rare.

Yet stochastic process models never died out. Despite their historic limitations in incorporating deterministic covariates, transaction timing models like the NBD remain useful as they are well suited to model recurring customer behavior in the absence of competitor data. As market researchers, we have the luxury of resources such as household panel data of competing brand purchases, but many companies and industries, e.g. financial services, lack access to similar household level competitor panel data. Most companies in most industries have to make do with the data they collect from their own customers. Nonetheless, this internal data can still be extremely valuable in addressing marketing research questions, as when Moe and Trusov (2011) utilize an exponential process to model how online product ratings affect transaction timing.

Before more thoroughly reviewing the early marketing stochastic process literature relevant to this thesis, I want to highlight two previous review articles and a book chapter. The first article, Morrison and Schmittlein (1988), is more a review of what was known about the NBD and some of its variants at that point in time, and less a literature review. They provide an excellent summary of the NBD itself, a distribution common to transaction timing models, and how it can be applied. The next two, Wagner and Tades (1987) and a book chapter by Fader, Hardie, and Sen (2014) have a more general focus on the stochastic process models that had been proposed for marketing research up through the 1980's. These two articles include models with higher order Markov switching probability matrices as well as the Linear Learning Model, both
of which make the assumption that past behavior can change at a later time, i.e. learning through usage, thus changing the rate of consumption for the individual. These models account for nonstationarity by allowing probabilities to change from one time period to the next, but like all stochastic process models they suffer from being difficult to parameterize in order to explain why.

Within this literature review, I will be mainly focused on zero-order stochastic process models that assume a steady rate of individual transactions, though there are a few notable exceptions that I will highlight as they come up. The zero-order rate can be adjusted by, say, seasonality or marketing activity that can be modeled through covariate parameters, but there is no probabilistic change to the underlying base rate. In the past, the zero-order assumption has proven to be robust. Bass, Givon, Kalwani, Reibstein, and Wright (1984) found that the zero-order assumption held with the majority of individuals in 10 different studies while using either stationary or nonstationary assumptions. There are situations where we would expect certain consumer behaviors to fundamentally change over time where a higher order model would be appropriate, such as after the introduction of a new product (Fader, Hardie, and Huang 2004). But a zero-order model provides a baseline model for when we do not expect the customer to fundamentally change their behavior at an instance in time.

In this chapter, I begin with a review motivating the continued application of stochastic process models. These models have sometimes been dismissed as mere curve-fitting exercises as they are often not parameterized on covariates. There are some important counterarguments that need to be considered before looking at the literature proper. Next, I review different transaction timing models. This will be mostly centered on the NBD and related variants that were proposed before research in stochastic process marketing models tapered off in the late-1980's. Once we reach the 1990’s, the marketing literature begins to explore the hazard function specification for transaction timing models. I will briefly explain the hazard function before reviewing that marketing literature. After the transaction hazard models, I consider the prominent zero-order
stochastic process choice models from that time period. This leads into the joint brand choice-purchase timing stochastic process models, some of which also adopt the hazard function as a non-utility-based approach to choice modeling.

**An Historical Defense of Stochastic Process Models**

The primary assumption behind applying stochastic models to marketing research is that the consumer behaviors we observe in the data can be modeled as if they were random. It then becomes the task of the researcher to fit the appropriate probability distribution to the data, or multiple distributions if multiple consumer processes are being modeled. This approach in many ways hearkens back to Karl Pearson. In the early marketing literature, researchers were generally clear they did not expect the observed behaviors to actually be random, as Goodhardt, Ehrenberg, and Chatfield (1984, p. 623, italics are the authors’) state:

This does not mean that we think a housewife times her purchases and chooses her brands literally at random (e.g. by tossing mental pennies). Instead, people generally have "verifiable reasons" for making a purchase (e.g. "We'd nearly run out", or "My mother-in-law was coming"). But taken *en masse*, purchases are capable of being modeled probabilistically at the individual level, even though true randomness lies only in the model.

While there is always a worry that omitted-variable biases can affect a model, stochastic models make the assumption that many of these deterministic causes, such as the mother-in-law coming to visit, will balance out so that the timing of observed behaviors are effectively random in the data. By itself, this is nothing new to statistical modeling; what is unobservable and unaccounted for in a model remains as an error term that is often referred to as the random effects component of a statistical model.

One of the constraints of stochastic process models prior to the mid-1980's is they were stationary models that for the most part did not include any explanatory covariates. The latent transaction timing rates in the models were often assumed to be invariant, while a basic
counterexample such as seasonal transactions shows this assumption to be false. But that by itself does not limit the usefulness of these models. As Greene (1982, p. 25, italics are the author's) pointed out using the NBD as an example:

In marketing, the stationary form of the [NBD] model assumes no change in the marketing environment, no change in the average rate of purchase or usage (no trend), and no change, even, in the rates of individual consumers (no shifting). The object of marketing efforts is to make these assumptions false, and the model is a powerful tool for detecting marketing efforts at the consumer level.

Prior to having a convenient form of modeling that could include explanatory covariates, stochastic process models were still able to provide an equilibrium baseline against which fluctuations in observed data could be compared.

Bass (1974, p. 2, italics are the author's) provides one of the stronger defenses for using a stochastic process model to help understand and describe customer behavior, particularly choice behavior:

If there is a stochastic element in the brain which influences choice, then it is not possible, even in principle, to predict or to understand completely the choice behavior of individual consumers. Moreover, even if behavior is caused but the bulk of the explanation lies in a multitude of variables which occur with unpredictable frequency, then, in practice, the process is stochastic. Such random occurrences as out-of-stock conditions, mother-in-law visits, and weather do influence choice behavior, but it is not possible or useful to include all such variables in empirical studies. It will never be possible to prove conclusively that behavior is fundamentally stochastic or fundamentally deterministic since it will never be possible to measure all of the variables which influence choice.

Stochastic process models make explicit the non-deterministic randomness found in the data, and they do so by fitting the data to a random distribution to describe what is going on. There is an implicit recognition in stochastic process models that we are not Laplace's demon. We do not have access to every possible causal explanation for what is observed in the data we collect. But behavioral patterns can still be reasonably fitted to a probabilistic distribution.

Logit and probit choice models, with their straightforward incorporation of covariates from which we can draw deterministic conclusions, replaced stochastic process models in marketing research for a good reason. It is the lack of covariates in many stochastic process
models that make them particularly vulnerable to the charge of being mere curve fitting exercises. However, it is important to understand the past value of stochastic process models as well as their motivating questions. Even if we treat consumer actions in data as if they were random and even though stochastic process models can lack explanatory covariates, stochastic process models are extremely capable of describing what is going on in a way that is relevant to managers as well as other professionals and practitioners. Stochastic process models can also provide a conditional expectation of what is likely to happen next that is well within reason. This may be less than satisfactory compared to a model that provides parameters fitted on deterministic covariates, as we rarely gain a theoretical explanation for observed behaviors in the data otherwise. But that said, most non-Bayesian models in use that lean on deterministic parameters for theoretical explanation are providing a mean of the behaviors observed, not a skewed spread of overall observed behaviors.

Wheat and Morrison (1990a) make two claims in favor of purchase choice (which they call purchase incidence) models over stochastic process models of intertransaction time. 1) Purchase choice models are robust to the possible discrete nature of intertransaction times, e.g., some customers shop on the same day of each week. 2) Purchase choice models are unaffected by right-censored data, and intertransaction time models are not. While these two points are true, they fail to address that we choose a model appropriate to the research question at hand, e.g. right-censored data is entirely relevant for applications such as whether a customer remains active or not. If we want to explore why a customer selects a brand, then yes, we will want some form of choice model. But if we are trying to understand why these transactions are timed the way they occur, which is certainly a managerially relevant question, then regardless of the added complexity a transaction timing model is the more appropriate model to use. Moreover, depending on the product category or industry, competitor choice data may not even be available to a firm, which renders moot which type of model to utilize.
Stochastic process models can provide a description of the overall distribution of behavior, whether skewed or even multi-modal. Once these models are extended with explanatory covariates we can ideally gain the benefit of modeling both the stochastic and deterministic elements together. That combination can help provide researchers and professionals a better explanation of why different patterns are being observed in the data, particularly when competitor data is unavailable. We can most easily understand the value of stochastic process models by starting with transaction timing models that are focused on modeling just one process.

*Transaction Timing Models*

The workhorse of purchase timing stochastic process models is the negative binomial distribution (NBD), and it was Ehrenberg (1959) who first introduced the NBD to marketing research. The NBD benefits from being an easy model to estimate, whether through maximum likelihood, method of moments, or even solving for the probability of no purchases against its expected value (Ehrenberg 1959, 1972). And as Ehrenberg (1972) demonstrates, the NBD is a powerful baseline model for describing stationary\(^1\) purchase patterns. Interestingly enough, even as early as 1959, Ehrenberg recognized that a stationary model would not generalize to all consumer buying behavior. Thus, he illustrated this claim with the example of a customer switching brand loyalty in the future, and there was also a need for nonstationary transaction timing models.

The underlying marketing assumptions for the NBD are that for a given item or brand, each consumer purchases with a mean rate that follows a Poisson distribution, which implies that

\(^1\) In probability theory, the terms homogeneous and nonhomogeneous are specifically used instead of stationary and nonstationary respectively for Poisson processes. These terms do have different, precise meanings in probability theory where stationary is a stricter term that requires exactly the same distribution while homogeneous is defined as occurring with the same rate. The marketing literature tends to conflate the terms stationary and homogeneous, with a usage preference for stationary. Hence, stationary and nonstationary are the terms used in this thesis.
the time between each transaction follows an exponential distribution (Ross 2010). These mean rates are then compounded with a Gamma distribution over all consumers, although Ehrenberg (1959) initially proposed a more specific chi-squared distribution. It was initially thought that the NBD could model the quantity purchased. Over time, the NBD was narrowed in focus to count the events of transaction incidence, i.e. the number of trips to the store when the transaction occurred over a period of time, since different package sizes and weights would confound what was being modeled (Chatfield and Goodhardt 1975).

Marketing researchers went on to use the NBD as a tool to effectively estimate transaction rates, market penetration, and more. While the NBD is a stationary model, counts of transaction incidence estimated from the NBD could be used as a baseline to compare against deviations in the data. Such deviations could then be attributed to changes due to marketing actions or seasonality, e.g. light users of a seasonal food would increase usage when it was in peak season while heavy users would not change their purchase behavior (Goodhardt and Ehrenberg 1967). Additionally, since transaction events could be treated as a Poisson process, this allows the NBD to model time between transaction events as well as event counts, providing additional descriptions of consumer behavior in the marketplace. The NBD was thus able to provide guidance to managers and researchers on the impacts of firm actions, and help them to recognize the behavior of different customer segments. And since the parameters of the NBD are easy to estimate from count data, the NBD is a tool that managers can use on their own.

The next advance for the NBD was to forecast what was likely to happen in the next time period based on the current time period. To solve for this conditional expectation, Goodhardt and Ehrenberg (1967) introduce the bivariate NBD that is modeled on two separate time periods. They use the term "conditional trend analysis" to examine deviations in purchase patterns from a stationary model, and they compare a stationary case with an unspecified food product that has a seasonal transaction pattern. It is important to note from their food example that the heavy users highlighted by Goodhardt and Ehrenberg (1967) do not change their usage with the season, so
seasonal transaction patterns may not necessarily be appropriately modeled by an overall population parameter. Morrison (1968) provides a Bayesian derivation of the same conditional expectation.

Goodhardt and Ehrenberg’s (1967) work highlighted the need to model customer segments with the NBD. Morrison (1969) introduces the zero-inflated NBD which jointly models two customer segments of active buyers and "hardcore" non-buyers. He demonstrates that without accounting for a group of non-buyers when using a stationary NBD, purchase expectations for light users will be underbiased and overbiased for heavy users. A zero-inflated model is applicable to data where zero values may be more frequent than would be predicted by the Poisson or NBD. This is typically seen in a sample that has a mode of 0 and another mode greater than 1. We can model these excess zeros such that we have one group of definite zeros with probability \( p \), and another group \( 1 - p \) distributed with probability mass distribution \( f(x) \) such that

\[
\begin{align*}
P(0) &= p + (1 - p)f(0) \\
P(x) &= (1 - p)f(x), x > 0
\end{align*}
\]

The term "zero-inflated" often used to describe this form of model originated with Lambert (1992) who was using a Poisson distribution for \( f(x) \), but Morrison (1969) is the first documented use that I have found. More recently, Venkatesan, Reinartz, and Ravishanker (2012) made use of the zero-inflated Poisson to model always-a-share purchase behavior with the excess zeros used to indicate a group of dormant customers. Note that \( f(0) \) is also the survival function, i.e. the probability the event has not happened. Within the marketing literature, when the survival function is used in a model likelihood for binary data, zero-inflated models are sometimes referred to as a split-hazard model (Sinha and Chandrashekar 1992).

Another challenge to the NBD is that purchase events can occur more regularly than would be predicted by a Poisson process model. Herniter (1971) noted this regularity, and suggests an Erlang-2 distribution for purchase timing rather than an exponential distribution to
account for this pattern. The Erlang-$k$ distributions are Gamma distributions with a fixed integer shape parameter $k$, with the exponential distribution included when $k$ equals 1. Herniter pointed out that if time between occurrences is distributed exponentially, it means the mode of the time interval is at zero, which is a questionable assumption. Chatfield and Goodhardt (1973) compounded the Erlang-2 distribution with a population Gamma distribution to create an equivalent to the NBD, which they call the Condensed NBD (CNBD), although in their empirical example the CNBD did not provide better fit compared to the NBD.

Schmittlein and Morrison (1983) develop the conditional distributions of the CNBD to forecast purchase events. They also pointed out the differences in the hazard rates of the NBD and CNBD that reflect the choice of exponential or Erlang-2, with the NBD having a high initial hazard rate reflecting a zero mode and the CNBD having a non-zero mode. The NBD is often the default assumption in modeling applications because it is linear in its expectation of future transactions conditional on current transaction data (Morrison and Schmittlein 1988), making the NBD an easily estimated baseline model against which we can compare more complex models, e.g., the CNBD.

While the NBD and its variants are useful tools to explain what is occurring in the marketplace, initially there was very little theory to explain why the NBD was an appropriate model to use. Schmittlein, Bemmaor, and Morrison (1985) tie the NBD to Beta-binomial and Dirichlet-multinomial brand choice models that I will discuss later in this chapter, but this was more to demonstrate how the NBD could follow from specific brand choice model assumptions. Kahn (1987) develops a theoretical model of intertransaction times based on usage rates rather than transaction timing. With specific distributional assumptions, her model does lead to the NBD, but her model also shows how other process distributions like the CNBD can occur as well. Her theory helps to clarify that a transaction event is based on a need, but the timing between the need and the observed transaction is different. For example, promotions can speed up when a transaction may occur, but it likely will not influence the underlying consumption of many
consumer goods such as, say, toilet paper or pet food. This will show up as increased variance without a change in the long-term transaction rate. Non-theoretical distributional assumptions are needed to arrive at either the NBD or CNBD, but Kahn gives us a theoretical statistical model that is tied closer to the individual consumer behavior.

What is needed to incorporate behavioral theories within the context of purchase timing models is the ability to parameterize the model on covariates. Gupta's (1991) article is a very important step in this regard, as he demonstrates a way to parameterize time-varying covariates with the Poisson and Condensed Poisson distributions as well as the compound NBD and CNBD models. Gupta’s method was to have a set of exponentiated covariates proportional to the distribution’s rate parameter. Extending the application of Gupta’s method to add covariates, Lenk, Rao, and Tibrewala (1993) propose a nonstationary NBD model with time varying covariates for predicting customer transactions from one period to the next with conditional trend analysis. Importantly, in Lenk et al. (1993)’s empirical analysis, the zero-class of customers remains under-predicted, which makes the case for a separate zero-class as well.

While Gupta's (1991) article came too late to have much influence on the overall direction of marketing models, it stands as an important reminder of why it can be worthwhile to go back to the older literature. Sometimes a problem has already been addressed, but forgotten. His article has been a critical piece in deriving the CLV modeling framework that I introduce in Chapter 5. All the same, these articles need to be understood in their original context, which can be difficult; even Fader, Hardie, and Sen (2014) mischaracterize Gupta's (1991) models as proportional hazard models, which is only true for the exponential case.

During this early time period in marketing science, there were a few distributions besides the NBD that were proposed for modeling transaction timing. Chatfield and Goodhardt (1970) propose the Beta-binomial distribution (BBD) as an alternative to the NBD, with time treated discretely. During each time unit (e.g. weeks) with probability $p$, a consumer will buy one or more times, where $p$ is a parameter of the binomial distribution over $n$ time units, and $p$ over all
consumers is distributed Beta. The NBD is a limiting case of the BBD, when \( n \) time periods go to infinity. While maximum likelihood and method of moments can be used to estimate the BBD parameters, Lee and Sabavala (1987) take a pre-MCMC Bayesian approach to the BBD, and, with simulated data, demonstrate that Bayesian inference can provide better estimates of the parameters.

Sabavala and Morrison (1981) note that the BBD was often used to model media penetration. However, there were times when the stationarity of the model was called into question. Sabavala and Morrison modified the BBD by adding a geometric process where the binomial \( p \) parameter can change after a media exposure occasion based on the rate of a geometric distribution, creating what they called the BBG model. This adds a nondeterministic, nonstationary element to the BBD. It is a model structure that should be noted as a potential extension for other NBD-like models, for example, Fader, Hardie, and Huang (2004) use a similar structure to allow a probability for the transaction rate to change from one transaction to the next.

Banerjee and Bhattacharyya (1976) propose that intertransaction times follow an inverse Gaussian distribution rather than an exponential or Erlang-\( k \) distribution. Since the inverse Gaussian is a two-parameter distribution, they use a bivariate Student's-\( t \) distribution for a compound distribution to model population heterogeneity. The fit for their model was not as good as the NBD in their empirical example. Sichel (1982) proposes the generalized inverse Gaussian distribution as a population mixing distribution for Poisson purchases rather than a Gamma distribution. As Sichel points out, the Gamma distribution is a special case of the generalized inverse Gaussian, but it is a three parameter distribution that has more flexibility. Whether this level of complexity is needed is an open question, but it is another example of how the heterogeneity assumption of the NBD can be modified.

Lawrence (1980) proposes a lognormal-lognormal distribution to replace the NBD where both individual transaction rates and population heterogeneity are distributed lognormal. However, his model had difficulty with the case of no purchase events. Additionally, it was
noted by Morrison (1981) that Lawrence was modeling individuals’ overall mean rates rather than time between transactions, with only population heterogeneity of those mean rates fitted to the data, indicating Lawrence misunderstood the nature of stochastic process transaction models.

Rather than a Gamma mixing distribution, Brockett, Golden, and Panjer (1996) maintain the assumption of a Poisson distribution for time between transactions, but then use a recursive set of probability distributions for compounding heterogeneity. This gives us their Generalized Compound Poisson-Pascal (GCPP) distribution, with the NBD as a specific instance of their model. In their empirical study, the zero-inflated GCPP outperforms the zero-inflated NBD.

Allenby, Leone, and Jen (1999) introduce what I would characterize as a generalization of the NBD. They model times between transactions with a generalized Gamma distribution. Since they used the scale version of this distribution rather than the rate, they model heterogeneity of the scale parameter with a generalized inverse Gamma distribution. Allenby et al. (1999) do not name their model, but it is certainly a more generalized Bayesian implementation of the NBD. They also note the usefulness of a stochastic process model when competitor data is lacking. They specifically state it is often not possible to estimate a construct like utility with company-only data. This is because the utility of a choice decision has no natural zero and must be referenced against another choice option in order to be quantified. In their application, they also introduce temporal dynamics with mass points for different segments and allow for segment switching. The downside of using a Gamma or generalized Gamma distribution to model intertransaction times is that it is difficult to include the survival function in the model to account for right-censored data. This is a point which Allenby and colleagues do not address. They also note that the generalized Gamma distribution can have a nonmonotonic hazard function, which will come up again in the next section when I review the marketing models that utilize a hazard function.

Late to this stream of literature modifying the NBD, Trinh et al. (2013) propose a Poisson distribution mixed with a lognormal distribution. These are not conjugate distributions, but with
modern computational power, their model can be readily estimated with simulation methods. In their empirical study with different product categories, their model predicts slightly better with multiple products in each category in the following year as compared to the NBD.

While the NBD and its variants have proven over time to be robust models, they are open to criticism. I have highlighted that they are challenging to parameterize on a set of time-varying covariates, though still possible as demonstrated by Gupta (1991). Additionally, NBD-like models do not have the direct appeal to theory that logit models have with their latent representation of utility that ties them to economic theory. However, as Allenby, Leone, and Jen (1999) note we cannot have a representation of utility if we have no competitor data to provide a reference against which we can quantify utility. Nevertheless, parameterizing transaction models was soon addressed through means of the hazard function.

**Hazard Function Transaction Models**

At the beginning of the 1990’s, the marketing literature began to utilize the hazard function of a stochastic process in order to model transaction data. Hazard models are designed to model time to an event. They are commonly used in the biostatistics literature to model expected time until death for causes such as cancer, heart disease, and other fatal medical conditions. The ability to parameterize hazard models on potential causes for an event occurrence makes them a useful tool for marketing problems dealing with time until an event, e.g. the next customer transaction.

The hazard function of a stochastic process is defined as follows. If we have a PDF of the probability of time until an event occurs, \( f(t) \), with its associated CDF \( F(t) \), and survival function \( 1 - F(t) \) (i.e., the probability the event has not yet happened), the instantaneous rate of an event occurring is the hazard function, and is defined as (Kalbfleisch and Prentice 2002):
\begin{equation}
(2.2) \quad \lambda(t) = \lim_{dt \to 0} \frac{P(t+dt \leq T < t+2dt | T \geq t)}{dt}.
\end{equation}

Taking the limit leads to the result
\begin{equation}
(2.3) \quad \lambda(t) = \frac{f(t)}{1-F(t)}.
\end{equation}

With a little algebraic manipulation, we can redefine the hazard function in terms of the survival function
\begin{equation}
(2.4) \quad \lambda(t) = \frac{f(t)}{1-F(t)} = -\frac{d}{dt} \log(1-F(t)).
\end{equation}

We can then redefine the survival function in terms of the hazard function
\begin{equation}
(2.5) \quad 1 - F(t) = \exp\left(-\int_{0}^{t} \lambda(u)du\right).
\end{equation}

From Equation 2.5, we can interpret the survival function of a stochastic process over time as a function of the summed hazards encountered during a period of time. Next, using Equations 2.3 and 2.5, we can also redefine the PDF solely in terms of the hazard function
\begin{equation}
(2.6) \quad f(t) = \lambda(t)[1 - F(t)] = \lambda(t)\exp\left(-\int_{0}^{t} \lambda(u)du\right).
\end{equation}

These are common statistical relationships for describing the probability of an event occurring in time, though these interrelationships are less appreciated in the marketing literature. The only distribution where the hazard function \(\lambda(t)\) is constant and invariant across time is the exponential distribution, \(f(t) = \lambda\exp(-\lambda t)\). Once we allow the PDF to vary with time, we need to ensure the model accounts for both the hazard function at time \(t\) as well as the integral in the survival function between our starting point in time and \(t\).

There are different methods for parameterizing the hazard function, but the method most commonly used is the proportional hazards model. This method is common because covariate parameters can be estimated semi-parametrically without reference to the underlying distribution.

To add covariates \(x\) proportional to the hazards function, we define:
\begin{equation}
(2.7) \quad \lambda(t) = \lambda_0(t)\exp(x'\beta),
\end{equation}
where $\lambda_0(t)$ is the baseline hazard function. This method of adding covariates as a multiplicatively effect on the hazard function is called a proportional hazards model in that we can compare the covariates that affect the lifetimes of, say, smokers and non-smokers, by modeling a ratio of different groups. This cancels out the baseline hazard function, leaving us with just the exponentiated covariates whose parameters can be estimated as a multinomial logit model. The hazard function must always be positive, hence the exponentiation of the covariates with their parameters.

Proportional hazards modeling is not the only method for adding covariates that could be utilized. One method from survival analysis for incorporating covariates, called accelerated failure time, is as a log-linear regression on the survival times (Lawless 2003). Another method is to add the exponentiated covariates rather than multiply them. But the proportional hazards method tends to be much easier to estimate. It is available in many statistical software packages and thus sees far more usage.

The first use in the marketing literature of a proportional hazards model with covariates that varied at each transaction comes from Jain and Vilcassim (1991) who model time until a transaction for coffee in their empirical study. Rather than an average consumption rate or household inventory, Jain and Vilcassim use the volume (in ounces or units) of the last purchase as a covariate. They tested five different transactions distributions: exponential, Erlang-2, Weibull, Gompertz, and a Box-Cox quadratic formulation of the hazard function. They found the Box-Cox quadratic, which has a nonmonotonic hazard function, fit their data best. Helsen and Schmittlein (1993) followed with a more thorough description of the importance of proportional hazards models to marketing researchers examining time duration data. In their empirical study of crackers, they found the Box-Cox quadratic distribution fit better than the Weibull. They were the first to incorporate a covariate for the price at transaction which had the expected negative parameter value. From economics, we expect a negative price coefficient as an increasing price implies lower consumer demand and fewer transactions over time. For a transaction timing
model, the implication of the negative price parameter is that consumers will delay a transaction when prices increase, and speed up a transaction when prices decrease.

There have been a few papers that followed which extended this initial research on transaction timing. Chintagunta and Haldar (1998) introduced a bivariate hazard function to examine the correlation of transaction timing behavior of two product categories. Their empirical study used the log-logistic distribution, which has a nonmonotonic hazard function. They found that the log-logistic performed better than the Box-Cox quadratic, inverse Gaussian, generalized Gamma, and semiparametric distributions. Seetharaman and Chintagunta (2003) is a more thorough investigation of different distributions with a proportional hazards specification to model purchase timing. Comparing the exponential, Erlang-2, Weibull, log-logistic, and expo-power distributions, they find the expo-power distribution (Saha and Hilton 1997), which also has a nonmonotonic hazard function, best fit their detergents and towels data (they did not test the Box-Cox quadratic). Their price coefficients were not always negative for some of the distributions they tested.

With an occasionally questionable price coefficient, the proportional hazards method may not always be the best method for adding covariates to a transaction timing model. Seetharaman (2004) proposes the use of an additive hazards model where

\[
\lambda(t) = \lambda_0(t) + \exp(x' \beta) .
\]

Seetharaman compares the additive hazards model with proportional hazards and accelerated failure time models when fitted to laundry detergent and paper towels transaction data. He finds the additive hazards model is more robust with different distributions compared to the exponential, Erlang-2, Weibull, log-logistic, and expo-power distributions. Using the additive hazards specification, the log-logistic distribution fit slightly better than the expo-power. Also, the proportional hazards model sometimes estimated positive price coefficients instead of the expected negative value, which the additive hazards and accelerated failure time models estimated successfully.
When multiple different distributions have been empirically tested, distributions with nonmonotonic hazard functions, such as the Box-Cox quadratic, log-logistic, expo-power, and the generalized Gamma (Allenby, Leone, and Jen 1999), tend to be the distributions that fit best transaction timing data. I note this as an opportunity that has not yet been fully researched in the CLV modeling literature which has primarily depended on the exponential distribution for the transaction process.

**Stochastic Brand Choice Models**

I include a review of the early choice stochastic choice literature in this thesis for two reasons. 1) Before the logit model swept up most modelers in its wake, choice models were an active segment of the stochastic process marketing model literature that was building more intricate models to answer more complex questions during the late 70's and early 80's. The first stochastic process CLV model, the Pareto/NBD of Schmittlein, Morrison, and Colombo (1987), was born out of the research questions generated during this fecund time period. Reviewing the choice models from this early time period can help us to understand stochastic process CLV models from their original context. 2) Many of the choice models from this time period used the NBD as a key building block. If we wish to pursue a research question such as how consumer choice affects customer value, we need a model capable of combining these areas of research. The stochastic process choice models from this time period provide us a means to do so. That is not an opportunity I will be pursuing in this thesis, but it does point towards how CLV research can be broadly extended to include choice decisions, something that has been initially investigated by Park, Park, and Schweidel (2014) and Schweidel, Park, and Jamal (2014).

Scanner panel data was not readily available until the early 1980's, but stochastic choice models were actively researched well before Guadagni and Little (1983) by using diary panel data and other sources. How a consumer goes about choosing an item for purchase amongst
competing items and why have always been key questions for marketing research. Models fitted to choice data go back to the early 1960's (Fader, Hardie, and Sen 2014). The historical stochastic process choice model literature has a multitude of models that are not zero-order and do specify mechanisms for preferences to alter over the observed time period being modeled. However, I will mainly be focused on zero-order choice models. As a stationary baseline, zero-order choice models provide a functional starting point to understand how these models were generally pieced together to describe the processes behind choice behavior.

One early model was the Beta-binomial distribution (Massy, Montgomery, and Morrison 1970) where the binomial distribution represented individual choice behavior that is compounded at the population level with a beta distribution. We have already seen the BBD used above as an equivalent model to the NBD (Chatfield and Goodhardt 1970), but where the time units are discrete. Many of the choice models proposed after the mid-1970s take advantage of how transaction timing can be related in a model to purchase choice within an overall joint model.

The BBD formed a solid starting point, but there is much more interest in understanding why a consumer chose to purchase something rather than when. This is reflected in the fact there was more interest in grounding the choice models that followed in theory. Herniter (1973) proposes a model based on the concept of entropy in that a system is in equilibrium at maximum entropy, i.e. at its most random state. Herniter’s theory borrows from the physical sciences for how an equilibrium system ought to behave and proposed that an equilibrium marketplace could be similarly modeled. Herniter’s model only needs market share data and aggregate consumer repeat and switch rates can be derived as outputs from the model.

Bass (1974) starts with Herniter’s (1973) theory to develop his own theory of stochastic preference. While Bass’s (1974) model is also aimed at predicting market share and switching probabilities in an equilibrium random state, he constructed his theory so that each brand has the NBD as a marginal distribution, in order to tie his theory to Ehrenberg’s (1972) work. Bass, Jeuland, and Wright (1976) expand on the theory from Bass (1974). Using the Luce formulation
for utility as a probability, which is then distributed Beta for population heterogeneity. They add a parameter to control the rate of brand switching. Their model provides a more solid statistical foundation for the theoretical model in Bass (1974). The purchase probabilities follow a Polya process, which is a mixed Poisson process. The model proposed by Bass and colleagues is one of the few instances I have found in the early stochastic process choice literature that does include a latent variable formulation for utility, so their model can be tied to economic theory.

Chatfield and Goodhardt (1975) propose a multivariate beta (Dirichlet) to model independent brand share, though they do not provide an empirical example. They then tie the underlying transaction rates to the NBD. It is the proportions of these rates that are fitted to a Dirichlet distribution. It is common for many of the stochastic process choice models that followed Chatfield and Goodhardt (1975) to establish a choice process derived from the transaction timing process. The independence assumption allows the Gamma distributions for each NBD for each brand to be summed up into a larger Gamma distribution. Shoemaker, Staelin, Kadane, and Shoaf (1977) follow Chatfield and Goodhardt (1975) by using a Dirichlet to model different brand purchase propensities. They do so with household level data to provide an empirical example. They also modify the parameters of the Dirichlet to allow for dependency based on purchase frequency, i.e., more frequent purchasers are likely to be more loyal.

Jeuland, Bass, and Wright (1980) combine Chatfield and Goodhardt’s (1975) model with Bass, Jeuland, and Wright’s (1976) model by creating a compound model with Gamma/Erlang-k brand transactions, Dirichlet-multinomial brand choice, and an additional parameter to control for brand switching. They describe this as a Multiple Hypergeometric model. What is worth noting here is how Jeuland and colleagues utilize a fuller realization of the hierarchical construction of different consumer processes into a larger, richer model.

Dalal, Lee, and Sabavala (1984) focus on an individual level variant of Chatfield and Goodhardt’s (1975). They propose a Poisson-Bernoulli model with the Poisson distribution for purchasing and Bernoulli for brand choice, but they allow the parameters to vary across the
population to allow for dependence between the two. Dalal and colleagues leave the population distribution unspecified, to focus on the dependence between the purchase and choice parameters. Schmittlein, Bemmaor, and Morrison (1985) describe the marginal Beta-binomial-NBD version of Jeuland, Bass, and Wright’s (1980) model and provide related conditional expectations. Schmittlein and colleagues focused on the marginal distribution as an application to describe whether a purchase is potentially captured in the diary data or not, back when scanner panel data was still not readily accessible.

Goodhardt, Ehrenberg, and Chatfield (1984) expand on their past work and proposed a composite model for brand choice and brand purchasing with a Dirichlet-multinomial for the brand choice process and NBD for brand quantity. In their model, they maintain independence across brands. They specifically note how their model follows from the limitations (p. 626), "In a strictly unsegmented market where the multinomial choice probabilities for each individual are fixed over time, the Dirichlet distribution is the only possible model." Their model was considered important enough to be given a full reading before the Royal Statistical Society and was published with commentary from other researchers. While similar to the model of Jeuland, Bass, and Wright (1980), Goodhardt and colleagues provide additional empirical evidence and different statistical characteristics for their model. Ehrenberg, Uncles, and Goodhardt (2004) forcefully argue the lawlike aspect of the NBD-Dirichlet model, at least for a steady-state market. While the characterization “lawlike” is questionable, their paper does provide working examples for how practitioners can apply the model to, for example, describe market dynamics.

What these models still lacked was the ability to incorporate explanatory covariates. The first direct parameterization on covariates that I have found in a zero-order stochastic process marketing model comes from Jones and Zufryden (1980, 1982) who propose a logit-NBD composite model to account for both choice and transaction timing. They used a logit transformation of choice to include covariates, where the dependent variable is distributed Beta for heterogeneity, which they choose in order to follow Chatfield and Goodhardt (1975) as the
Beta distribution is the marginal distribution of the Dirichlet. While they only looked at one brand, they were able to demonstrate a demographic effect on choice behavior alongside quantity purchased. Following economic theory, the logit model specifically has a theoretical extreme value distribution (McFadden 1974) rather than a Beta distribution. However, the beta distribution is very flexible and is likely to be close enough to allow for an easier to estimate joint model with the NBD.

Wrigley and Dunn (1985) demonstrate how to incorporate independent variables into either the NBD or the Dirichlet-multinomial, though not together as a joint model. Their approach was in the spirit of generalized linear modeling. Wagner and Taudes (1986) propose a joint NBD-Dirichlet model to accommodate nonstationary explanatory variables. Wagner and Taudes start at the individual level with a multivariate Poisson process, i.e. a Polya process, over multiple brands. The multinomial captures the probability of a brand being chosen, and Wagner and Taudes use a latent variable formulation of utility that captures the explanatory variables from the Poisson process. Heterogeneity is introduced with a multivariate Gamma distribution for purchases and a Dirichlet distribution for population choice. What is truly significant here is the introduction of a nonstationary transaction process with seasonality, retail price, and advertising budgets brought in as covariates on the transaction rate. Wagner and Taudes (1991) continue their research by extending their model with customer specific nonstationary variables, and they also discuss how their model compares to the NBD-Dirichlet of Goodhardt, Ehrenberg, and Chatfield (1984). The 1986 and 1991 articles by Wagner and Taudes are complex, but they are an important demonstration of what is possible by combining different process models.

Gupta's (1991) work to include time-varying covariates in NBD and CNBD models rests heavily on the covariate formulation first used by Wagner and Taudes (1986).

Gupta (1988) proposes a model that separately estimated intertransaction times (Erlang-2), brand choice (logit), and quantity (cumulative logit). Gupta argues his data reflected independence, so each model could be separately estimated. All three parts of his model included
covariates, and none of these three processes are mixed at the population level. But what his model proposes, and the independence assumption had been quite common through this time period, is a way to deterministically examine each customer process to provide a more complete picture of the marketplace.

In the 1990’s, as marketing academics applied hazard modeling to their research questions, we have a few examples of a hazard model being utilized in choice modeling. Vilcassim and Jain (1991) use multiple different hazard functions for each household to represent timing the change from one brand transaction to another. For example, one hazard model measures the timing to purchase of the same brand while additional hazard models are utilized for the timing to switch to other brands in the product category. Their models include time-varying covariates.

Wedel, Kamakura, DeSarbo and ter Hofstede (1995) start with Vilcassim and Jain (1991) and add segment heterogeneity. In their empirical study of ketchup, they say they improve upon Kamakura and Russell (1989) for predicting holdout data without making use of utility theory in their model, though I note Wedel and colleagues only model two segments whereas Kamakura and Russell (1989) modeled five segments. Popkowski, Leszczyc, and Bass (1998) is another paper that begins with Vilcassim and Jain (1991), but they expand their model into continuous time. One potential extension for these hazard-choice models could be by means of a Markov switching matrix in order to embed the choice decision into the matrix probabilities, and separate choice decisions from the timing of the next transaction, while still including both the brand choice decisions and transaction timing information in the model, but that is outside the scope of this thesis.
Non-Transaction Stochastic Process Marketing Models

At the end of the 1990’s as computational power improved, we begin to see stochastic process and hazard models applied to events of interest to marketers besides transactions and brand choice. Additionally, more complex models are proposed during this time period by combining more than one stochastic process to explain more complex data that looks at consumer events of interest besides transaction timing and brand choice. Fader, Hardie, and Lee’s (2005a) BG/NBD reintroduced this line of research to CLV models. And most of this multi-process research since then has been focused mainly on CLV models.

Churn, i.e. when a customer leaves, is also an important event for marketing researchers to model and forecast. Using survey data, Bolton (1998) uses a proportional hazards model of the length of duration of cellular service, and how the duration length is affected by customer satisfaction. Schweidel, Fader, and Bradlow (2008a) also applied a proportional hazard model to cellular service retention. Their model includes time-varying covariates that do not depend on survey data and so has broader application. They also found the Weibull distribution fit better than the exponential. In another application of survey data to examine causes of churn, Braun and Schweidel (2011) use a competing risk hazard model that, with telecom survey data where previous customers have provided reasons for leaving, allows them to measure churn that is either controllable or uncontrollable to the firm.

Also worth noting as a model of an event of interest besides transactions or churn, Park and Fader (2004) utilize a bivariate stochastic timing model to examine the correlation between browsing two different web sites. Schweidel, Fader, and Bradlow (2008b) use a bivariate hazard model to jointly examine acquisition and retention of HBO service. They found the log-logistic fit better than the Weibull or expo-power distributions. And Moe and Trusov (2011) apply an exponential hazard model with time-varying covariates to model when users will post online reviews and how posting affects future transactions.
While not a CLV model, the RFM (recency of last purchase, frequency of transacting, and cumulative monetary value) targeting model proposed by Colombo and Jiang (1999) has received attention from the CLV modeling literature, mainly because of its customer spending model. Colombo and Jiang model transaction frequency using the NBD. They introduce the Gamma/Gamma compound stochastic process model to estimate customer spending. Frequency and spending were assumed to be independent and estimated separately. The Gamma/Gamma model assumes that each consumer may spend an amount that follows a Gamma distribution. Population heterogeneity is estimated by allowing the individual Gamma distribution's rate parameter for each customer to be distributed Gamma over all customers. This is analogous to how heterogeneity is introduced to the Poisson distribution in order to create the NBD. Their model allows for great flexibility regardless of the underlying data since the Gamma distribution can model many different shapes. The Gamma/Gamma model went on to be used to model consumer spending in a CLV applications by Fader, Hardie, and Lee (2005b) as well as Glady, Baesens, and Croux (2009).

In marketing, we expect consumer usage behavior to change between acquisition and repeated usage as they become familiar with a product or service. Fader, Hardie, and Huang (2004) propose a stochastic model that uses a changepoint formulation to allow for the transaction rate parameter to change with time from the initial purchase. They found the exponential distribution fit better than the Erlang-2. Schweidel and Fader (2009) modify Fader, Hardie, and Huang (2004) so there is only one state change between trial and regular usage. During the trial period, the time between transactions is distributed exponential, which then switches to Erlang-2 when repeat customers achieve a steady usage state. Their model only slightly improves upon Fader, Hardie, and Huang (2004). It may have benefitted from a zero-inflated specification to model non-repeat customers for further improvement.

Jen, Chou, and Allenby (2009) point out that not estimating interdependence between quantity and timing can lead to biased estimates. This is potentially correctable with a covariate
of the amount last purchased, e.g. Jain and Vilcassim (1991). The most important part of their model is transforming it into a conditional model to allow for simpler inverse Gamma priors on the variance rather than an inverted Wishart distribution on a larger covariance matrix. They use a log transformation of quantity over time, which can be rewritten as log(quantity) – log(time). This linear form allows for separate priors. They then model log(quantity) conditional on log(time). In a direct marketing setting, Jen and colleagues empirically found modeling dependence between quantity purchased and intertransaction durations improved prediction of both quantity and intertransaction times.

The base of nearly all of these examples from the marketing stochastic process model literature is a transaction timing model, usually the NBD or a very close variant. We have seen in the literature how a transaction timing model can be extended by accounting for a choice process as well. The customer lifetime value model literature makes a different extension to transaction process models by adding a second stochastic process to model how long we expect a customer to remain a customer. CLV models are the focus of the next chapter.
CHAPTER 3: STOCHASTIC PROCESS CLV MODELS

The analytic structure for estimating customer lifetime value (CLV) is generally a time value of money calculation that gives us the net present value of expected future profits adjusted by a discount rate (Gupta et al. 2006). Taking a typical present value formula for a stream of constant cash flows, with constant retention rate $r$, constant discount rate $d$, and letting time go to infinity, this is simply the expected recurring margin for each time period modified by a multiplier composed of the retention and discount rates.

$$CLV = \sum_{t=0}^{\infty} \frac{(Revenue_t - Costs_t) r}{(1+d)^t} = Margin \sum_{t=0}^{\infty} \frac{r}{(1+d)^t} = Margin \frac{r}{(1+d-r)}, \ n \to \infty$$

There is a conceptual problem in treating the expected future revenue stream from a customer as we would, say, a corporate bond, as customers are most definitely not emotionless, inanimate, fixed-income financial instruments, regardless of whether they are contractually obliged or not. The idea that customers would continue to have the same margins and retention rates over an extended period of time has no face validity. But if we wish to estimate the value a customer is expected to provide the firm in the future, net present value does provide a useful structure as long as we are cognizant of its limitations. For example, we can use a much shorter, finite time frame where we know most observed customer patterns in the data we are analyzing are relatively stable, though this does call into question the "lifetime" terminology that is often used (Fader and Hardie 2014). Adding restrictions to account for the limits of how well we understand customers allows us to utilize a net present value framework while still having confidence in the output from CLV models.

Most CLV models have focused on estimating two components of the future value calculation, mainly because these two components are customer behaviors that are difficult to model. The first component is the probability a customer is still active, often referred to as $P(Alive)$, while the second is the expected number of future transactions from customers, which is
conditional on the customer remaining active. In the sparse instances where customer spending has been addressed in CLV models, researchers recognize that spending is dependent on the number of transactions so in most of these models the spending process of the model does little more than account for a regression to the mean level of overall spending per transaction, e.g. Schmittlein and Peterson (1994). Fewer models account for the discount rate and, where they do, generally these models just utilize a constant discount rate. It is understandable that less effort has been focused on spending and discount rates; how often a customer is willing to transact, and until when, are more difficult values to model. On the other hand, what a customer will spend over multiple future transactions is very close to their average past spending per transaction. Additionally, since managers have some control over the costs per transaction (and thus their overall margins), future transactions and lifetimes have important managerial value that might be lost by a model's focus on future revenue. This thesis follows most of the existing CLV modeling literature and does not include the monetary component of CLV in its models.

A major portion of the marketing literature for CLV models is devoted to stochastic process CLV models, beginning with the Pareto/NBD (Schmittlein, Morrison, and Colombo 1987) and continuing more recently with variants like the BG/NBD (Fader, Hardie, and Lee 2005a). What makes these stochastic CLV models valuable is they are able to provide forward-looking estimates of purchase patterns and customer churn when the event of a customer leaving the firm is unobservable which is typical for most noncontractual consumer goods and services. In these settings, stochastic process models still allow us to model a probability for whether a customer remains active. To understand the underlying assumptions behind these CLV models, we need to understand how the model assumptions fit with observed transaction data, which I address in the next section of this chapter.

After assessing how CLV model likelihood functions are generally structured for the individual customer, I will discuss the limitations of these models with regards to nonstationary data as well as with customers that have no observed repeat transactions. Both kinds of data lead
to biased model estimates. This bias is intrinsic to the individual model likelihood functions used by any stochastic process CLV model similar to the Pareto/NBD. That leads to a discussion of why adding time-varying covariates to CLV models has been a challenge. Since time-varying covariates are common in other settings, such as choice modeling, quite a few quantitative marketing researchers have posed to me various forms of the question, "How is adding time-varying covariates even a problem?" CLV models need to account for what might be happening in the time period between the last observed purchase and the end of the observation period. Accounting for this right-censored data requires statistical techniques that are common in survival modeling, but not as well understood outside of that domain. Modeling right-censored data is where mistakes have been made in the marketing literature, not just with CLV models. I will wrap up the chapter with a review of the relevant CLV modeling literature.

**The Generalized Structure of Stationary Stochastic Process CLV Models**

Below in Figure 3.1 is an example of what we might observe with non-contractual customer purchase data, where each X represents a transaction took place in that time period. Customer A has transacted in the most recent time period, so it is highly likely they are still active as a customer. Customers B and C have not transacted in some time, represented by the right-censored grey area. Customer C has been an infrequent purchaser, so it seems intuitive they may still be active. With Customer B, we see a customer who had been transacting frequently, but has then stopped for a much longer period than they had normally transacted, so intuitively B is more likely to no longer be an active customer.
Figure 3.1: Example of Observed Purchase Data

<table>
<thead>
<tr>
<th>Time Period</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer A</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Customer B</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Customer C</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Stochastic process CLV models make use of transaction process models like the exponential distribution (which implies a Poisson count distribution) to model observed transaction behavior, but CLV models also make use of the right-censored period of non-activity under observation to estimate whether or not a customer remains active. Along with a transaction process, CLV models add an additional, latent stochastic process to account for that period of inactivity, which may be permanent. This is referred to as the latent attrition process. Stochastic models like the Pareto/NBD (Schmittlein, Morrison, and Colombo 1987) and the BG/NBD (Fader, Hardie, and Lee 2005a) make the assumption that a customer lost has truly left the firm. There are other CLV models which assume customers can go dormant rather than go away permanently, e.g. Venkatesan and Kumar (2004). This assumption is sometimes referred to as "always-a-share" and I will address those models in the literature review later in this chapter. There is not an a priori reason to assume whether a customer may be either dormant or truly lost for good. However, one of these assumptions must be chosen a posteriori by the researcher, depending on the data and the transaction context, in order to model when a customer stops being a customer.

The basic structure that is used for modeling customer attrition is similar to the composition of the NBD; one distribution is utilized for purposes of modeling the individual action. A second distribution is used as a compound, or prior, distribution to model customer heterogeneity. The relevant data that is used is the censored time period between the last observed purchase and the end of the observation period. Two events could be happening during the right-censored time where we observe no transactions by the customer: 1) the customer left,
and we use the attrition process to model the probability of leaving, or 2) the customer is active
but has not yet transacted, in which case we use the survival function of the transaction process,
i.e. the probability of 0 transactions since the last observed transaction. The observed transaction
history thus indirectly affects which event, either attrition or transaction, is more likely after the
last observed transaction, depending on how much time has passed. Effectively, a stochastic
process CLV model estimates how likely a customer has left by means of a probabilistic race
between the attrition and transaction processes and which is likely to happen first. If a customer
only transacts once a year, they are unlikely to have left six months after their most recent
purchase whereas a customer that transacts approximately monthly is much more likely to have
left six months after their most recent transaction. Stochastic process CLV models take this
customer heterogeneity into account when assigning a probability of whether a customer is still
active or not.

The model likelihood of the observed data of an individual customer is constructed from
the transaction and attrition processes, which I describe below. However, this model likelihood
only represents what has already been observed. Fitting the model does not by itself produce the
estimates of a customer’s future transactions nor the probability of whether a customer remains
active. These two estimates are created by using the model parameters estimated from model
fitting. The estimated model parameters are themselves subject to bias when the model is fitted
to the observed data.

I describe here in a general manner how the individual likelihood function is constructed
with the transaction and attrition stochastic processes. Later in this section when I discuss how
the model parameters and estimates can be biased during model fitting, I explain that these biases
are intrinsic to a generalized but uncorrected likelihood function. The likelihood function of a
CLV model is composed as follows:
(3.2) \[ L(t) = \text{ (likelihood of observed purchases by the customer) } \times \]
\[ \text{ (likelihood the customer survives through those purchases) } \times \]
\[ \text{ (likelihood the customer is either active through or left during the right-censored time period).} \]

We then model the likelihood for an individual customer with two stochastic processes as

\[
(3.3) \quad L(n, t_n, T|f(\cdot), g(\cdot)) = \prod_{j=1}^{n} \left[ f(t_j)(1 - G(t_j)) \right] \times \\
\left\{ (1 - F(T - t_n))(1 - G(T - t_n)) + \int_{t_n}^{T} [1 - F(u)]g(u)du \right\},
\]

where \( f(\cdot), F(\cdot), \) and \( 1-F(\cdot) \) are the PDF, CDF, and survival function respectively of the transaction process, \( g(\cdot), G(\cdot), \) and \( 1-G(\cdot) \) are the PDF, CDF, and survival function respectively of the latent attrition process, \( n \) is the number of observed repeat transactions, \( T \) is the total observed time, \( t_j \) indicates the time between each repeat transaction \( j \) from 1 to \( n \), and \( t_n \) is the time of the last observed transaction. The product term in the first set of brackets represents the repeat transactions we have observed, where for the attrition process we assume the customer has remained active through those repeat transactions. The second set of brackets models the right-censored time period between the last transaction and the end of the observation period. In this generalized example, the transaction and attrition processes are assumed to be independent in that they are multiplied together without an explicit covariance modeled, but there is conditional dependence implicit in the likelihood through \( n \) in that each customer has remained active through their observed transactions.

In the articles from the CLV literature where a covariance is explicitly modeled, e.g. through a bivariate lognormal prior (Abe 2009), it is rare that adding an explicit covariance provides any noticeable improvement to estimates of future purchases even when the covariance parameter is significant. I explore this further in Chapter 6 where I show that a bivariate lognormal/logitnormal prior fitted to simulated datasets with independent transaction and attrition processes will estimate a significant, nontrivial, positive correlation between the two processes.
without improvements to estimates of future behavior. Correctly modeling the implicit conditional dependence between the two processes remains unsolved in the CLV modeling literature.

For models that assume a customer can return after a period of dormancy, we do need a stochastic process that does more than just terminate the customer. In those cases a hidden Markov model is often used to incorporate a switching matrix that accommodates the change between one state and the other. But even with these more complex models, the conceptual structure of the CLV model at the individual level remains the same.

When both the transaction and attrition processes are modeled with continuous distributions (and \( u \) is unknown since we cannot observe when a customer leaves), the integral over the censored time period that integrates out \( u \) from the two processes highlights where latent attrition models can become problematic if we wish to change the model assumptions away from stationary distributions. To simplify the individual likelihood, Fader, Hardie, and Lee (2005a) propose using a geometric distribution for the attrition process. They assume that if a customer left, it happens right after a transaction. Their use of the geometric distribution maintains a memoryless distribution to model when a customer leaves, similar to the Pareto/NBD using an exponential distribution for the individual attrition process, only as a discrete distribution instead of a continuous one. Using a discrete distribution for the attrition process eliminates the need for an individual level integral in the likelihood, since if the customer left we assume the event happened at time \( t_{n} \), while still providing nearly the same estimates as the Pareto/NBD. This simplifying assumption allows us to focus additional model complexity through the transaction process, as our observed data is often associated with just the transactional data. Assuming a geometric attrition process with parameter \( p \) which indexes at 0 as proposed by Batislam, Denizel, and Filiztekin (2007) for their MBG/NBD model in order to account for customers with zero repeat transactions, the individual likelihood becomes
\begin{equation}
L(T, n|f(t), p) = \left[ \prod_{j=1}^{n} f(t_j) \right] (1 - p)^n \times \left[ p + (1 - p)(1 - F(T - t_n)) \right];
\end{equation}

\[ n = 0, 1, 2, ... . \]

I will discuss later in this chapter that further modeling steps are needed to correctly account for customers with no repeat transactions.

With some notable exceptions, the stochastic process CLV model literature has maintained the original assumption of the Pareto/NBD that the number of customer transactions (distributed $f(t)$ in Equations 3.3 and 3.4), follows a stationary Poisson purchase process. This implies that over time periods of equal length, say, the last four weeks and the next four weeks, a customer is expected to purchase roughly the same quantity in each time period. Formally, a stationary Poisson process implies that the time between each purchase follows an exponential distribution (Ross 2010). Other transaction distributions can certainly be assumed. For example, the CLV literature has yet to explore various distributions such as the log-logistic distribution mentioned in Chapter 2. However, the underlying conceptual structure of Equations 3.3 and 3.4 will hold for any stochastic process CLV model with non-contractual data that assumes a lost customer is lost for good. In Chapter 5, I will extend Equation 3.4 into a generalized CLV modeling framework that includes time-varying covariates and a zero-inflated model specification.

Why Time-Varying Covariates Are Not Common in the CLV Literature

As I mentioned earlier in this chapter, quite a few quantitative marketing researchers have asked me why adding time-varying covariates is even a problem. This has never been an issue for most econometric applications, such as choice modeling. It comes down to how CLV models must account for the right-censored time period between the last observed purchase and the end of the observation time period. In particular, it is necessary to incorporate the survival function of the observed transaction process within the model likelihood in order to account for when there
are no more observations in the right-censored time period. That additional survival function complicates how we incorporate time-varying covariates into a CLV model.

With most choice models and game theory structural models, the right-censored time period between the last observed purchase and the end of the observation period is never under consideration. There is a basic assumption in these models that any patterns in the observed data will continue to hold after the last observed transaction of each consumer. There is no need to include statistical instruments like survival functions in the model likelihood. That assumption greatly simplifies how these models need to account for covariates that vary over time.

Survival modelers are fully justified in asking why adding time-varying covariates is a problem since many of the methods for including time-varying covariates while accounting for right-censored data come from their literature. Survival modelers know what to do because they have spent enough time on the problem to understand how to handle time-varying covariates correctly. But survival modelers would also pose the question, "Which method should be used for adding covariates?" That there are multiple methods to choose from in the survival analysis literature ought to be indication that adding time-varying covariates to a nonstationary stochastic process should be handled with due consideration and care. The proportional hazards method that I utilize in Chapter 5 to add covariates into my CLV modeling framework is more common in the marketing literature, but proper attention to the details remains necessary.

CLV models bring an additional set of challenges by combining at least two stochastic processes that need to have their survival functions included in the model likelihood. If we look back to Equation 3.3, the integral over \( u \) in the right-censored time period is particularly problematic once we add time-varying covariates into the mix of either or both transaction and attrition processes. If both stochastic process distributions are continuous, both the integral and the deterministic parameters on the time-varying covariates of the rate parameters need to be solved. In such cases, it is questionable whether we have a tractable problem. Making the
assumption that the attrition process is discrete simplifies the likelihood, e.g. Equation 3.4, which then makes it considerably easier to incorporate time-varying covariates.

Some of the authors in the CLV literature appear implicitly aware of the issue by avoiding time-varying covariates. Fader and Hardie (2007) provide analytic derivations for how to include time-invariant covariates in the Pareto/NBD and the BG/NBD. Unfortunately, to date time-varying covariates have not been included in their models. As I have found, incorporating time-varying covariates into CLV models is genuinely a difficult modeling problem. Jerath, Fader, and Hardie (2011) even go so far in their conclusion as to question whether there would be any benefit to be gained from the added model complexity required to add parameterized covariates to a CLV model. In this thesis, I show that seasonality leads to biased estimates. So it is worth the effort to include time-varying covariates.

The Uncorrected Individual Likelihood Function Leads to Biased Parameter and Model Estimates

In this section, I describe how CLV models can produce biased estimates of future customer activity when they are fitted to nonstationary transaction data that includes customers with no repeat transactions. The structure of the individual likelihood function described in Equations 3.3 and 3.4 is responsible for this bias. The CLV modeling framework I develop in Chapter 5 corrects for these two issues, but there will remain an open question with regards to what is the “best” transaction distribution to use.

Seasonality Produces Biased CLV Model Estimates

Seasonality affects stationary latent attrition models through the right-censored time period of length \( T - t_n \). By Slutsky’s Theorem (Goldberger 1991), the length of the right-
censored time period, \( T - t_n \), converges in distribution when \( t_n \) is independently and identically distributed (i.i.d., which also implies the distribution is stationary). Consequently, when \( t_n \) is i.i.d., latent attrition model estimates will be consistent and unbiased. But when transactions are more likely to occur during specific periods of time (e.g., during the holidays), the time between \( t_n \) and \( T \) will be more clustered together within the population rather than randomly distributed because \( t_n \) is no longer i.i.d. Thus, seasonal transaction data affects the inference of the model parameters because the length of the right-censored time period is no longer randomly distributed.

However, there is a more direct effect on CLV model estimates through \( P(Alive) \), the probability the customer remains active at the end of the observation period, which is estimated after fitting the model likelihood. \( P(Alive) \) for a customer is defined as the individual likelihood, conditional on that customer being active at the end of the observation period, proportional to the full individual likelihood,

\[
P(Alive) = \frac{\text{[Probability of no transaction between } t_n \text{ and } T \mid \text{customer active}}}{\text{[Probability of no transaction between } t_n \text{ and } T]}\text{,}
\]

which using Equation 3.3 becomes

\[
P(Alive) = \frac{\prod_{j=1}^{n}[f(t_j)[1-G(t_j)]\times[(1-F(T-t_n))(1-G(T-t_n))]}}{\prod_{j=1}^{n}[f(t_j)[1-G(t_j)]\times[(1-F(T-t_n))(1-G(T-t_n))]+\int_{t_n}^{T}(1-F(t))g(t)dt]}
\]

when we assume a continuous attrition process, whereas with Equation 3.4

\[
P(Alive) = \frac{\left[\prod_{j=1}^{n}f(t_j)[(1-p)^n]\times[(1-p)(1-F(T-t_n))]\right]}{\left[\prod_{j=1}^{n}f(t_j)[(1-p)^n]\times[p+(1-p)(1-F(T-t_n))]\right]}, n = 0, 1, 2, ...
\]

when we assume a geometric distribution for the attrition process. In both forms, the product term in the likelihood that represents the observed repeat transactions cancels out from the numerator and denominator. Thus, \( P(Alive) \) is solely estimated from the right-censored portion of the individual likelihood. As the observed time since the last transaction, \( T - t_n \), decreases or increases, the numerator of \( P(Alive) \) becomes larger or smaller relative to the denominator as the survival functions of both processes, \( 1 - F(\cdot) \) and \( 1 - G(\cdot) \), get closer to 1 or 0 respectively.
As a result, estimates of \( P(Alive) \) will be smaller or larger based on when within a seasonal cycle the model is estimated. And since latent attrition model estimates of future transactions are generated by multiplying an expected future transaction count by \( P(Alive) \) to account for the probability the customer may have already left, the stationary transaction assumption leads to biased estimates of future transactions for a population that exhibits nonstationary transaction patterns primarily through \( P(Alive) \). Estimates of future transactions right after a peak transaction period will overestimate future transactions because the time between \( t_n \) and \( T \) will be small for most of the population, whereas estimates right after an extended seasonal drought in transaction activity will underestimate future transactions. The seasonal effect on both \( P(Alive) \) and estimates of future transactions is demonstrated in the simulation and empirical studies in Chapter 5.

**Zero-Repurchasing Customers Bias Parameter Estimates**

The structure of CLV models also allows customers with no repeat transactions to bias model parameters and the resulting model estimates. When customers have no repeat transactions, the first part of the individual likelihood in Equations 3.3 or 3.4 representing observed repeat transactions becomes 1, with \( t_n = 0 \). We are left with the portion of the individual likelihood representing the right-censored time period. I focus here on the likelihood with the geometric latent attrition process: \([p + (1 - p)(1 - F(T - t_n))]\) from Equation 3.4.

This likelihood for an individual with zero repeat transactions can be maximized in two ways, either \( p = 1 \) or \( F(T) = 0 \) when \( 0 \leq p < 1 \). Neither mixing distribution for population heterogeneity of the transaction and attrition processes will allow for these extreme points to be reached. But what happens during model inference is the transaction population distribution lets \( F(T) \) be estimated closer to 0 while \( p \) is estimated closer to 1, which biases the overall inference of the model parameters. There is a likelihood maximization tradeoff here between \( p \) and \( F(T) \).
that is impossible to isolate analytically. What I have found using simulated data is that \( p \) is often far more affected than \( F(T) \), but not always, and the exceptions seem to be dependent on the transaction distribution being used.

Using a zero-inflated model specification corrects for this bias as it effectively isolates a set of customers with no repeat transactions into their own latent class. I will discuss how to apply the zero-inflation model specification in Chapter 5. Additionally, the simulation study in Chapter 5 demonstrates the bias with inference of the geometric attrition process in an uncorrected CLV model, with increased estimates of \( E[p] \) well above the true value, which is corrected for by using the zero-inflated specification.

Which Transaction Distribution Is “Best?”

There remains an open question as to which transaction distribution is best, at least with regards to the current structure of the individual model likelihood described by Equations 3.3 and 3.4. With fixed shape parameters, e.g. with the exponential or Erlang-2 distributions, I demonstrate with simulated data in Chapter 5 that with a seasonal correction and a zero-inflated model specification that we improve estimates of the rate parameter, though with some trouble identifying the correct transaction distribution by means of model estimates related to the transaction process. Once there is an additional transaction process parameter besides the rate parameter to estimate, it becomes more difficult to recover the shape parameter successfully. This is particularly noticeable with the Weibull shape parameter, as I will show with fitting my modeling framework to simulated data in Chapter 6.

I find this to be troubling. As I noted in the last chapter, the marketing literature has generally favored transaction distributions with nonmonotonic hazard functions. The challenge during model fitting is how to separate estimation of the transaction process in the portion of the likelihood function that covers the right-censored time period from joint estimation with the
attrition process. I described above the right-censored part of the individual likelihood function is susceptible to bias from customers with no repeat transactions, and it likewise presents an ongoing challenge to identify the best transaction distribution for the data being analyzed. This question will not be addressed in this thesis, but it does make for an opportunity for future research.

**CLV Model Likelihood Discussion**

I have made use of generalized distributions for the individual transaction and attrition processes within this section to highlight two issues with the assumptions underlying CLV models, that of nonstationary seasonal data and that of customers with no repeat transactions. This in no way dismisses what latent attrition CLV models have achieved. We are attempting to model transactions up until a customer leaves, an act that happens only once by model assumption, and is unobservable that single instance it does happen. That leaves us with a vast amount of uncertainty to untangle. Current CLV models do provide reasonable forecasts of future customer transactions (Wübben and von Wangenheim 2008) despite the inherent uncertainty, and that is impressive. But if we want to provide managers less biased model estimates, and many managerial decisions require estimates that precise, the two major obstacles that we have highlighted in the individual model likelihood remain. In Chapter 5, the modeling framework I develop for CLV models remedies the two issues of nonstationary seasonal transaction patterns and an excess of customers with no repeat transactions.

**Stochastic Process CLV Model Literature Review**

With the first stochastic process CLV model, the Pareto/NBD, Schmittlein, Morrison, and Colombo (1987) start with the NBD to represent the overall customer transaction process, and
then added an individual exponential process to model the time when a customer leaves, compounded with a Gamma distribution to model customer heterogeneity, which together form a Pareto distribution of the second kind. Schmittlein and colleagues also made the assumption that the transaction and attrition processes are independent. This independence assumption is often maintained in stochastic CLV models, mainly for ease of deriving the overall likelihood formula for purposes of maximum likelihood estimation, though there is still dependence implicit in the model likelihood. More recent CLV models that utilize hierarchical Bayesian inference, e.g. Abe (2009), do explicitly model dependence between the transaction and attrition processes, though in Chapter 6 I will call into question how well we can correctly model this dependence.

Altogether, six modeling assumptions underlie nearly every stochastic process CLV model that extends the likelihood function of Equation 3.3, and more specifically the latent attrition models that are similar to the Pareto/NBD:

1) A customer that leaves is either lost for good, or not.

2) Each customer transaction process is distributed with a customer specific rate parameter, e.g. Poisson distributed counts or exponentially distributed intertransaction times.

3) Heterogeneity amongst the customer transaction rate parameters is modeled with a compound, or prior, distribution, e.g. a Gamma distribution compounded with a Poisson distribution will form the NBD for the overall population transaction process.

4) Each customer attrition process is distributed with a customer specific rate parameter, e.g. exponentially distributed.

5) Heterogeneity amongst the customer attrition rate parameters is modeled with a compound, or prior, distribution, e.g. a Gamma distribution compounded with an exponential distribution will form the Pareto distribution for the overall population death process.

6) Customer transaction and attrition processes are either assumed to be independent, or have their dependence explicitly modeled.
The CLV literature has extended these models by modifying various distributional assumptions, usually the attrition process, and sometimes removing the independence assumption as well. But the overall customer/population hierarchical structure, with separate transaction and attrition processes, has remained in place. Schmittlein and colleagues acknowledged that their Pareto/NBD was incomplete. They proposed that time-invariant covariates could be incorporated into the model, although an actual method was not developed until Fader and Hardie (2007). And Schmittlein and colleagues recognized that the Gamma compound distribution assumptions could be relaxed. But because the Pareto/NBD is unwieldy to estimate, usage of the model has been somewhat limited.

The CLV literature that followed Schmittlein, Morrison, and Colomb (1987) can be roughly divided into three overlapping streams. The first stream takes a given CLV model as tools to be applied in an empirical setting and sometimes extends a model with an additional process such as money spent or a discount rate. This stream of the literature has been more focused on answering strategic managerial questions. The second stream is devoted to developing new CLV models, and began in the marketing literature with an always-a-share model proposed by Venkatesan and Kumar (2004). This modeling stream accelerated in 2005 when Fader, Hardie, and Lee (2005a) published their BG/NBD model. Most of the stochastic process CLV models published since then have been similar variants of the original Pareto/NBD. Some of these CLV models do use state change stochastic processes rather than simple attrition or transaction processes. I highlight those models separately. The third modeling stream is more recent and seeks to expand the scope of how CLV models can be applied, e.g. extended with customer choice decisions. The models coming out of this third stream are more computationally complex.

There is a large marketing strategy literature dedicated to CLV, sometimes under the terminology of customer equity (Blattberg and Deighton 1996) or customer asset management (Bolton, Lemon, and Verhoef 2004). These studies focus on the strategic implications of CLV
for the firm. There is also an extensive modeling literature devoted to service and customer relationship management in general, to which Rust and Chung (2006) provide an excellent review. The focus of this thesis is more specifically on stochastic process CLV models that are latent attrition models, but I note for the reader that the overall CLV and CRM literatures are much broader in scope.

Applied and Slightly Modified Stochastic Process CLV Models

The Pareto/NBD was not empirically validated until Schmittlein and Peterson (1994), who fit the Pareto/NBD in a business-to-business setting. Schmittlein and Peterson also added an independent monetary component through a regression to the mean of a normal distribution. Reinartz and Kumar (2000) utilize the Pareto/NBD with consumer catalog data to challenge the long-held assumption that the relationship between customer lifetime and customer profitability is positive, though for their empirical data this is tied to the mailing decision costs to the firm. That is, firms can overmarket to a customer to the point where the firm has spent more on marketing costs than the profitability of that customer. Reinartz and Kumar (2003) extend their analysis by turning their estimates of the probability that the customer is still active into lengths of profitable time. They then used a proportional hazards model of those profitable lifetimes to estimate what customer and marketing variables lead to longer profitable customer lifetimes. This use of a hazard model would not work to estimate CLV with existing non-contractual customers, but it does demonstrate how the estimates from stochastic process models fitted on past customers can be further analyzed to provide managerially relevant information.

Fader, Hardie, and Lee (2005b) combine the Pareto/NBD with an independent Gamma/Gamma spending model from Colombo and Jiang (1999) to estimate CLV. Fader and colleagues put the Pareto/NBD within an RFM (recency of last purchase, frequency of transactions, and cumulative monetary value) context by comparing their model estimates with
recency and frequency. They also create a Discounted Expected Transactions (DET) metric to add a constant discount rate to their CLV calculations.

Wübben and von Wangenheim (2008) compare the Pareto/NBD and the BG/NBD to a set of managerial heuristics. They found that the Pareto/NBD and BG/NBD perform better in predicting future transactions in terms of RMSE, but they do a poor job of predicting whether a customer is still active compared to a basic managerial heuristic of declaring a customer dead after a long enough period of inactivity. I will revisit this in Chapter 5, as the poor prediction of a customer leaving is correctable with a zero-inflated model specification.

Glady, Baesens, and Croux (2009) take the Pareto/NBD, add profit with a Gamma/Gamma submodel following Fader, Hardie, and Lee (2005b), but they allow profit to be dependent on the number of transactions. Through means of a regression coefficient Glady and colleagues tie together the Pareto/NBD with the Gamma/Gamma profit submodel. Their model worked well with banking data, but did not show improvement with the CDNOW data used by Fader, Hardie, and Lee (2005b).

Fader and Hardie (2010) add customer segmentation to the modeling of CLV. They introduce a model they call discounted expected residual lifetime (DERL), an extension of the BG/NBD, that provides a probabilistic expectation of how long a customer is likely to remain (adjusted by a discount rate). Because DERL depends on how many transactions through which the customer has survived, it likely needs a fair amount of customer activity to calibrate appropriately. But Fader and Hardie make two important qualitative points regarding CLV. First, CLV should be treated as a probabilistic expectation, i.e. $E(\text{CLV})$ rather than a deterministic measure, which is a point that may be more intended for the marketing strategy audience. Second, customer segmentation is needed to avoid biasing estimates of CLV, which they demonstrate by accounting for how different cohorts of customers change over time in terms of DERL as some customers leave and others stay. That said, because latent attrition CLV models
account for individual probabilities rather than cohort summaries, and Fader and Hardie reference their BG/NBD, these models may not necessarily need additional segmentation schemes.

Zhang, Bradlow, and Small (2015) provide another view on interpreting the output from CLV models by introducing what they call clumpiness as a value to examine alongside RFM. Clumpiness means that customers can go through transaction streaks. Applying the BG/BB model introduced by Fader, Hardie, and Shang (2010), Zhang and colleagues use a separate measure for clumpiness to show that purchasing in streaks is an indicator that a group of customers is more likely to be alive and have more monetary value. By itself, their application demonstrates how nonstationary customer behavior can affect customer value. Chapter 5 will address nonstationary transactions within the CLV model itself.

Finally, there have been a handful of recently proposed minor modifications to existing CLV models. Hoppe and Wagner (2014) modify the Pareto/NBD by extending the observed data the model can calibrate on beyond transactions, e.g. customer activity such as calling customer service. This can create a smaller right-censored time period than between the time of the last observed purchase and the end of the observation period. Using empirical data from a prepaid wireless service where there is no contract, they demonstrate better estimates that customers were still active compared to the Pareto/NBD. Karvanen, Rantanen, and Luoma (2014) incorporate survey data into a CLV model. With wireless phone customers, they calculate repurchase intention from customers’ self-reported past transaction history to estimate the equity of the group by brand customer is using. And Marshall (2015) develops a form of nonparametric estimation of the Pareto/NBD. Marshall ignores having population distributions, but the transaction and attrition processes are based on the individual level data. His proposal did not perform as well for out of sample future estimates with two datasets, CDNOW and a retail store in Turkey, used for validation by other CLV models.
New Stochastic Process CLV Models After the Pareto/NBD

Beginning in the mid-2000's, as interest in CRM applications increased, so interest in metrics like CLV increased. Most of the stochastic process CLV models from this time period follow the lost-for-good assumption of the Pareto/NBD, but not all. The first new CLV model after the Pareto/NBD that I have found comes from outside of the marketing literature. Wu and Chen (2000) propose separately estimated models of independent components of CLV. They use a logit model to estimate whether a customer is a one-time purchaser, a linear regression to determine the number of repeat purchases, and then an Erlang-$k$ model with Gamma heterogeneity to determine the time between those repeat transactions. Wu and Chen published in a management science journal, and the regression for the number of repeat transactions is an unusual choice for a CLV model, which is likely why their paper rarely cited. But I believe their model is the first CLV models to suggest Erlang-$k$ distributions to model the time between transactions. Fader and Hardie (2002) provide a response to Wu and Chen with a zero-inflated NBD with a shifted index to account for one transaction, but no repeat transactions. Their simple model predicts almost as well as Wu and Chen’s.

Venkatesan and Kumar (2004) propose the first new stochastic process CLV models within the marketing literature after the Pareto/NBD. They assume a customer might be dormant for a period, but they have no attrition process. Their empirical data is a business-to-business environment where it is less likely the lost-for-good assumption will hold. Following Allenby, Leone, and Jen (1999), they used a generalized Gamma distribution to model transaction frequency. They modify the transaction rate by subgroup membership based on covariates. They state the survival function of the generalized Gamma was used in time periods where there was right-censoring in the purchases, indicating dormancy. Venkatesan and Kumar were not clear how this was implemented for parameter inference, which is problematic as the survival function of the generalized Gamma distribution does not have a closed form.
Venkatesan, Kumar, and Bohling (2007) extend Venkatesan and Kumar (2004) by adding time-varying covariates to the purchase rate parameter. This is first CLV model with time-varying covariates, though specifically for an always-a-share model. However, when looking at the web appendix of their paper, the MCMC sampler described does not include the survival function. This omission raises concerns about their 2004 article.

Most of the CLV models proposed after 2004 maintain the NBD assumption of the Pareto/NBD with individual transactions distributed exponentially rather than something more complex like the generalized Gamma of Allenby, Leone, and Jen (1999). Historically, the NBD has been a robust distributional assumption for modeling purchase behavior that has proven itself in multiple contexts (Morrison and Schmittlein 1988). There is comparatively less clarity in the CLV literature on the best way to model the probability of whether a customer has left. However, keep in mind that with the lost-for-good assumption we model an event that is assumed to occur only once and is unobservable when it does occur. The first assumption change proposed with regards to the attrition process comes from Fader, Hardie, and Lee (2005a) who replaced the Pareto lifetime portion of the Pareto/NBD with a Beta-geometric distribution. The two assumptions of the beta-geometric distribution are 1) that after any purchase a customer will become inactive with probability $p$, and 2) that heterogeneity in $p$ follows a Beta distribution. Fader and colleagues found that expected future transactions estimated with their BG/NBD correlated 0.996 with the Pareto/NBD when analyzing the same data set. This indicates that little is lost by using a model with a simpler attrition process. However, a drawback of the BG/NBD is it assumes customers with no repeat transactions remain active. That is a questionable assumption. It leads to the BG/NBD performing worse at estimating inactive customers as compared with other CLV models (e.g. Jerath, Fader, and Hardie 2011). Fader and Hardie (2007) demonstrate how to incorporate time-invariant covariates in the BG/NBD as well as the Pareto/NBD.
The formulation of the Beta-geometric distribution Fader and colleagues used, which assumes there is at least one more transaction after the first, is also known as the shifted Beta-geometric. In later work, the term sBG/NBD is sometimes used to specifically refer to the Fader, Hardie, and Lee (2005a) model. One drawback of this formulation, which Fader and colleagues acknowledge, is that with data where active transactions are infrequent the BG/NBD does not fit as well to the data as the Pareto/NBD. Batislam, Denizel, and Filiztekin (2007) and Hoppe and Wagner (2007) independently developed a variant of the BG/NBD that changes the assumption a customer will become inactive with probability $p$ after any purchase to allow the possibility of attrition after the very first transaction. Batislam and colleagues called their model the MBG/NBD, which stands for modified BG/NBD.

There are other attrition process variations proposed over the last few years that are also worth noting. Jerath, Fader, and Hardie (2011) propose a model they call Periodic Death Opportunity (PDO), which changes the Beta-geometric assumption a customer will become inactive with probability $p$ after any purchase to be after any time interval instead. They then demonstrate when the length of the time interval approaches zero that the Pareto/NBD is the limiting case. One other result they pointed out is that different CLV models predict different lifetime lengths, with the BG/NBD having poor results in their example.

Bemmaor and Glad (2012) replace the assumption that customer lifetimes are exponentially distributed with a Gompertz distribution, but maintain the Gamma mixture to form their G/G/NBD model. While the Gompertz distribution is different from the Gamma distribution, it is similar to the Weibull distribution in having an exponentially increasing failure rate. Both the Gompertz and the Weibull distributions are more commonly used in actuarial applications as they simplify modeling accelerating or decelerating death rates. All of the CLV models I have described that are variants of the Pareto/NBD but with modified attrition process assumptions provide very similar fits to empirical data. Some models fit specific data better than
others. This is not surprising with the lost-for-good assumption where we are modeling an event that occurs once and is unobservable when it does happen.

Modifying the transaction distribution assumption is potentially more promising, though less common, since transactions are often the only observed data we have available for non-contractual goods and services. Platzer's (2008) model from his unpublished Master's thesis is particularly worth noting as he directly challenges the standard Poisson process assumption in stochastic CLV models. Platzer starts with the MBG/NBD, and modifies it by assuming the interpurchase times are distributed Erlang-\(k\), i.e. Gamma distributed with a fixed integer shape parameter \(k\) (the NBD case is included when \(k\) equals 1). His proposed model is called the MBG/CNBD-\(k\), and with his model Platzer placed second in a CLV competition hosted by the Direct Marketing Education Foundation in 2008. Platzer and Reutterer (2016) utilize Gamma distributions for the individual transaction processes rather than the Erlang-\(k\) (their Pareto/GGG model) which further improves CLV model estimates compared to the NBD assumption.

Another proposal to modify the purchase process comes from Fader, Hardie, and Shang (2010) who begin with the BG/NBD but assume that transactions follow a binomial distribution that is compounded with a Beta distribution to account for customer heterogeneity. This replaces the NBD with a Beta-binomial distribution (Chatfield and Goodhardt 1970 show the NBD is the limiting case of the Beta-binomial), with the resulting model called the BG/BB. Fader and colleagues suggest the BG/BB be used for modeling larger discrete time intervals, e.g. yearly. With time units that large, a simpler transaction distribution than, say, Erlang-\(k\), makes sense. They also propose a different mixing distribution by taking logit transformations of the geometric and binomial parameters, and then giving the logits of the two parameters a bivariate normal distribution. This formulation would allow adding covariates through the normal distribution, though Fader and colleagues do not take advantage of this. Using Bayesian inference with the bivariate normal prior, their empirical data did have a correlation of 0.36 between the transaction
and attrition processes rate parameters. This approach only slightly improved the model fit and did not noticeably improve prediction.

The use of Bayesian inference has allowed for the estimation of more complicated stochastic process CLV models. Borle, Singh, and Jain (2008) demonstrate the flexibility of a Bayesian approach with data from a purchasing membership club where the act of leaving the club is observed. They combined a NBD for transaction behavior, a discrete hazard model for lifetime estimation, and a lognormal distribution for purchase amounts. All three distribution rates were allowed to covary with each other although improvement was mixed compared to an independent model structure. Their use of the NBD on individual data, rather than a Poisson-Gamma mixture, is unusual. Their data also had time of exit information, which allowed for the use of a hazard model to estimate lifetimes independently. I note that every individual in their sample of 1,000 individuals had an exit time, which could bias their model estimates by not including customers that are still active. With no censored data, no survival function was needed. It effectively means that CLV estimates from their model are based on completely dead people and are backwards looking. They did include time-varying covariates, particularly a quadratic time series correction, but the lack of a transaction survival function is questionable.

Singh, Borle, and Jain (2009) generalize their 2008 work. Once again, they propose the direct draw of individual rates from an NBD or Pareto distribution in their MCMC iterations rather than the underlying Poisson or exponential distributions with the usual Gamma priors for population heterogeneity. They even go so far as to propose a Gamma/Gamma transaction/attrition model. While this may provide a decent approximation, it lacks the expected hierarchical structure of a stochastic process CLV model. And while stating that their approach is generalizable, they fail to address the right-censored time period in any way as once again they fail to include the survival function of the transaction and attrition processes.

Abe (2009) proposes the first true hierarchical Bayes variant of the Pareto/NBD. Abe utilizes a bivariate lognormal rather than gammas for the prior distribution. This prior allows for
regression directly on the means of the purchase and death parameters, as well as allows for
dependence between the two. His empirical applications included time-invariant covariates in
their regressions. Korkmaz, Fok, and Kuik (2014) have a working paper of a latent class
extension of the Abe (2009) model that also allows for time-invariant covariates to sort
individuals into one class or another. Korkmaz and colleagues use a Bayesian mixture of normals
formulation to model different groups.

Some empirical situations will require models that are more complex than what I have
described so far. This is particularly so when we expect a state change in a process that
significantly alters the rate from one time period to the next. There are a few specific examples
from the literature. For example, Ho, Park, and Zhou (2006) modify the Pareto/NBD by allowing
both the transaction and attrition rates to transition after a previous transaction. They use a
hidden Markov transition matrix to handle the state change, though their model is fitted to
simulated data, not actual data. Their model is an interesting inclusion of nonstationarity to the
Pareto/NBD by incorporating a state change. Ho and colleagues propose that a change in
satisfaction would be the reason for the change in state, but there are no behavioral measures, or
any covariates, included in their model. While satisfaction is a possibility for a state change, so
are many other changes in household attitudes and behaviors.

Netzer, Lattin, and Srinivasan (2008) use a hidden Markov model in a CRM application
to model people in different states of engagement, in their case college alumni. The segmentation
captures propensity to donate depending on state, and what helps to migrate alumni to an engaged
state. Ascarza and Hardie (2013) develop a similar model that jointly models usage and churn in
a contractual setting. Both usage by customers and eventual churn are predicted by a latent
variable of commitment, which is modeled using a hidden Markov model that allows for different
states of commitment. Potentially, some form of multi-state hidden Markov model could be used
for Pareto/NBD style models if we want a latent state besides whether the customer has left.
Ma and Büschken (2011) modify the BG/NBD to add a process for customers to return from dormancy. Once a customer is inactive, their model adds an independent exponential time distribution for when the customer returns rather than a Markov transition matrix. If a customer does not come back, or at least not within the observation time period, their model is effectively the BG/NBD. Over short time periods, this also allows for customers to have a purchase rate and a time of inactivity, which could be one way to address seasonality. This added process hearkens back to Sabavala and Morrison (1981) and their BBG model as well as Fader, Hardie, and Huang (2004), though Ma and Büschken do not reference either model. Ma, Tan, and Shu (2015) expand on this concept with the Pareto/NBD to model when customers should be contacted with marketing to try to reactivate them, though in their empirical study they do not correct for the endogeneity of the contact data.

More Complex Stochastic Process CLV Models

Within the last few years, we have started to see more complex and innovative stochastic process CLV models that seek to extend which customer behaviors are to be modeled as components of CLV. Schweidel and Knox (2013) have the first stochastic process CLV model with time-varying covariates under the lost-for-good assumption. They extend what affects CLV by including direct marketing contacts. I have some issues with their approach, but they took care of a lot of details. First, the main difference between their approach and the Pareto/NBD type models is that Schweidel and Knox use a Bernoulli transaction process and they did not generalize to other transaction process distributions. Second, they do not have the hierarchical population mixture distributions that typify models that utilize and extend the NBD, such as the Pareto/NBD. Now that said, one interesting contribution is their use of a copula\(^2\) to correlate

---

\(^2\) For more information about copulas, see Danaher and Smith (2011). Copulas are a statistical method for constructing multivariate distributions from marginal distributions, and are useful in
purchase probability with the amount donated, where the copula implicitly acts as the population mixing distribution. However, they model attrition as a hidden Markov model, tying it to both timing and amount donated, though there is no population mixing distribution for this attrition process. The use of amount donated is also odd since they use a DMEF donation dataset where the correlation between the number of donations and the amount donated is near 0. As part of their model, they do add both a Winter dummy variable and RFM variables for the firm’s nonrandom mail decision. They also use another copula to model how the firm’s decision to mail the customer is tied to the customer’s decision to purchase. This copula is meant to follow the endogeneity correction proposed by Park and Gupta (2012), but I do not see in Schweidel and Knox (2013) the nonparametric error correction that Park and Gupta describe. Schweidel and Knox also incorporate latent classes by allowing for the population level regression and copula parameters to be different for each class, though these are fixed parameters rather than hierarchical population mixing distributions.

Park, Park, and Schweidel (2014) extend the BG/NBD into multiple categories at once. In their model, the transaction process becomes a timing model for arriving at a store. They do have time-varying covariates on the arrival process, but I have a question with regards to their survival function (included here as Equation 3.8 where \( i \) is an individual and \( iJ \) is the last store arrival of that individual) as it is not a summation of the hazard function over the right-censored time period as discussed in Chapter 2:

\[
(3.8) \quad 1 - F(T) = \exp \left( -\lambda_{ij}(T - t_{ij}) \right).
\]

Effectively, this formulation of the survival function states that the covariates last encountered in the retail setting, which are contained in \( \lambda_{ij} \), remain active while a customer is not in the store. There may be specific reasons for this assumption, but it should be stated explicitly. Park and colleagues mention in a footnote that they tried other distribution assumptions like the Erlang-2 situations where we need to model the correlation between completely different distributions, such as a continuous and a discrete distribution.
and the Weibull for the arrival process, but found no significant improvement in their model. I suspect their formulation of the survival function may have had an impact on their conclusion, though my simulation study in Chapter 5 will show that the incorrect transactional distribution can have a better fit statistic than the correct transactional distribution. That said, Park and colleagues make a very important extension to the CLV literature by adding a shopping basket choice process across categories from both Russell and Peterson (2000) and Moon and Russell (2008) to the timing process. Despite my questioning of the timing process survival function, this is an important first step in tying CLV models to choice models so we can begin to analyze how one influences the other.

Schweidel, Park, and Jamal (2014) take a different approach to incorporate choice decisions along with a component of CLV. Schweidel and colleagues start with a bivariate choice model between two options, purchasing digital content and posting digital content. They then add a geometric latent attrition process that includes time-varying covariates to capture activity until the customer leaves, while their transaction process is implicitly binomial.

van Oest and Knox (2011) modify the MBG/NBD to add a complaint process. Complaints are modeled as a separate Poisson process. Once a complaint arrives, it changes the transaction and attrition rates. This incorporates an increased risk of attrition after a customer complains. Knox and van Oest (2014) expand on their previous model with a few additions. They allow for transactions to follow the Weibull distribution, and they handle the specification of the survival function appropriately for the right-censored time period. I note they use a discrete mixture of latent classes to model population heterogeneity rather than a hierarchical distribution that typifies the CLV modeling literature. But I describe in Chapter 6 that I have had issues recovering the model parameters when using a Weibull distribution for transactions, and eliminating the population mixing distributions may be one way around this problem. I also note they provide no information on recovering the model parameters by means of simulation. Knox and van Oest also parameterize the transaction and attrition processes on a distribution parameter.
that varies with each transaction and complaint event, and they use lagged-by-event variables to account for endogeneity.

Gladys, Lemmens, and Croux (2015) start with the Pareto/NBD, add a Gamma/Gamma distribution for monetary value, and add time-invariant covariates on the rate parameters for the three processes. Their innovation is the use of copulas over all three population processes, and a fourth copula at the individual level between transaction timing and purchase amounts. Finally, Braun, Schweidel, and Stein (2015) extend the BG/NBD so that the \( p \) parameter is parameterized with covariates, though not with any covariates that need to be corrected for endogeneity. In their empirical study, they examined how a mismatch between a requested level of service and the delivered service affected retention.

The CLV modeling literature has become an increasingly active branch of the academic marketing literature since 2005. CLV is clearly an important metric for managers. As I mentioned in the introductory chapter, CLV is a key driver of the firm’s value to shareholders (Gupta, Lehmann, and Stuart 2004; Gupta et al. 2006; Kumar and Shah 2009). And CLV has been incorporated into many managerial resource allocation decisions such as service quality, advertising, direct marketing, and distribution (Bolton, Lemon, and Verhoef 2004; Kumar et al. 2008). But more model methodological work is needed in order to provide managers with better estimates of the transaction and attrition components of CLV, such as a generalized method of adding time-varying covariates.

**CLV Models with Time-Varying Covariates and Endogeneity Correction**

I have found seven research articles in total that include time-varying covariates within stochastic process CLV models, which are described in Table 3.1. Two of these articles are relevant to this thesis. Schweidel and Knox (2013) add time-varying covariates through a Bernoulli transaction process. And Park, Park, and Schweidel (2014) misspecify the survival
function of a nonstationary exponential transaction process, which as I have discussed should be interpreted that incorporating time-varying covariates is not straightforward, and should not otherwise detract from an article that expands a CLV model across product categories. Only one of these articles, Schweidel and Knox (2013) incorporates a covariate to adjust for seasonal transaction rates, without exploring how much nonstationary seasonal data can bias the estimates from CLV model. I note that the Schweidel and Knox empirical study had a time unit granularity of three months, i.e. quarterly. Additionally, none of these seven articles proposes a generalizable method for adding time-varying covariates to stochastic process CLV models. That primary contribution of this thesis will be developed as a modeling framework in Chapter 5.

Table 3.1: CLV Articles with Time-Varying Covariates

<table>
<thead>
<tr>
<th>Article</th>
<th>Seasonal Correction?</th>
<th>Endogeneity Correction?</th>
<th>Empirical Setting</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Park, Park, and Schweidel (2014)</td>
<td>No</td>
<td>No</td>
<td>Department store beauty care products</td>
<td>Open questions regarding survival function</td>
</tr>
<tr>
<td>Knox and van Oest (2014)</td>
<td>No</td>
<td>Lagged variables</td>
<td>Internet and catalog retailer</td>
<td>Dummy variable for complaint made after last transaction event. Latent classes for population heterogeneity.</td>
</tr>
<tr>
<td>Braun, Schweidel, and Stein (2015)</td>
<td>No</td>
<td>No</td>
<td>Freelance writing services</td>
<td>Only attrition process is parameterized with covariates</td>
</tr>
</tbody>
</table>
Similarly, correcting for the endogeneity inherent in marketing variables that could affect CLV, endogenous because the underlying managerial decisions are nonrandom, is similarly rare in the CLV modeling literature. Much of the data used in CLV models comes from direct marketing data where the marketing targeting decisions are based on RFM (recency of last transaction, frequency of transactions, monetary value) customer behaviors. This creates a difficult structural problem in that in order to account for how marketing affects customer value, we need to model both the managerial decision to target customers based on their purchasing behaviors while also modeling how those customer behaviors affect customer value. Hence, most published CLV models forego the inclusion of marketing variables as independent variables. Along those lines, the two most productive authors in the CLV modeling literature, Peter Fader and Bruce Hardie, have even questioned the basic benefit of adding covariates to CLV models (Jerath, Fader, and Hardie 2011).

Of the articles I have reviewed above, five address endogeneity. Venkatesan and Kumar (2004) and Venkatesan, Kumar, and Bohling (2007) use a lagged-variables autocorrelation method for endogeneity correction proposed by Villas-Boas and Winer (1999). Villas-Boas and Winer proposed their method to correct for endogenous pricing which is not normally set at an individual customer level based on that customer's behavior, so it is unclear whether this method of endogeneity correction would be appropriate to correct for the endogeneity of RFM targeting. Venkatesan, Reinartz, and Ravishanker (2012), whose CLV model does not include an attrition process, use a structural method proposed by Manchanda, Rossi, and Chintagunta (2004), to correct for the endogeneity of marketing contacts. All three of these CLV models are in business-to-business settings where the customer is assumed to sometimes go dormant but never die, and in the case of Venkatesan, Reinartz, and Ravishanker (2012) they use similar pharmaceutical marketing data as that used by Manchanda and colleagues. Knox and van Oest (2014) correct for the endogenous managerial decision to intervene with a customer based on whether the customer
complained in the previous time period; in other words, the decision is modeled as a reaction to a
customer action by use of lagged variables. The last example of endogeneity correction in a CLV
model comes from Schweidel and Knox (2013) who apply the copula endogeneity correction

The structural method of endogeneity correction of Manchanda and colleagues does
demonstrate a pattern in many structural endogeneity correction models in that the equations of
the model are of a similar form. With Manchanda et al. (2004) the equations are counting
processes with the NBD used to model the number of prescriptions written by a doctor and the
Poisson distribution used to model the number of sales call to that doctor. The structural method
of endogeneity correction proposed by Manchanda and colleagues, where they are accounting for
marketing contacts by pharmaceutical representatives to doctors in order to measure the effect of
those contacts on prescription writing, would be far more difficult to implement in a stochastic
process CLV model. This is due to the survival functions in the right-censored time period within
the individual likelihood functions. In the individual CLV likelihood function, we have both the
CDF (as the survival function) and its derivative, the PDF, for the transaction process, which
means we have a differential equation. So we would be looking at a system of two differential
equations to model both the managerial decision and the customer response in order to account
for the right-censored time period and how both customer and managerial decisions change over
time. This form of endogeneity correction in a CLV model would be similar in mathematical
structure to a predator-prey population size model. That is far outside the scope of this thesis.

This structure of like equations is also common in the direct marketing literature that
incorporates an endogeneity correction. For example, Donkers, Paap, Jonker, and Franses (2006),
Van Diepen, Donkers, and Franses (2009), and Rhee and Russell (2009) all use choice models to
structurally account for both the firm's decision to mail someone and the separate consumer
decision to respond to that mailing. Van Diepen and colleagues even construct their covariates in
a way to represent changing dynamics over time. While structural models in general do not have
to be of like form in terms of their equations, it does certainly simplify how one equation informs the other equations in a structural model. The copula endogeneity correction method of Park and Gupta (2012) does at least avoid the issue of attempting to construct structural models of like equations, which could be of potential benefit to the researcher depending on the modeling context. All that said, the inclusion of marketing variables in CLV models remains rare because of the dynamic structural nature of the endogeneity correction needed for the direct marketing contact data often used in existing CLV models.
CHAPTER 4: DIFFERENT MODEL ASSUMPTIONS AND DATASET DESCRIPTION

I take a brief digression with this chapter. First, I discuss the different models that I will be using in the empirical and simulation studies in Chapter 5. Next, I describe the empirical dataset that will be used in the empirical study. Development of the CLV modeling framework in the next chapter is generalized, but the studies do make use of this chapter.

**Different Model Assumptions**

In the analyses in the following chapter, I will be applying my CLV modeling framework, developed in Chapter 5, as an extension of Equation 3.4 above where \( f(t) \), i.e. the distribution of the time between transactions, is one of either the exponential, Erlang-2, or Erlang-3 distributions. While the Weibull distribution had also been proposed for this thesis, there are issues using the Weibull distribution that I document in Chapter 6. For the attrition process, I assume a geometric distribution that begins indexing at 0, per Equation 3.4. For comparative purposes with simulated data, I also use a Bayesian formulation of the Pareto/NBD (Schmittlein, Morrison, and Colombo 1987) as a specific form of Equation 3.3 described below.

Equations 3.3 and 3.4 are repeated here for convenience.

\[
(3.3) \quad L(n, t_n, T|f(\cdot), g(\cdot)) = \prod_{j=1}^{n} \left[f(t_j)\left[1 - G(t_j)\right]\right] \times \\
\quad \left[\{1 - F(T - t_n)\} \{1 - G(T - t_n)\} + \int_{t_n}^{T} \{1 - F(u)\} g(u) du\right],
\]

\[
(3.4) \quad L(n, t_n, T|f(\cdot), p) = \left[\prod_{j=1}^{n} f(t_j)\right](1 - p)^n \times [p + (1 - p)\{1 - F(T - t_n)\}] ,
\]

where \( f(\cdot), F(\cdot), \) and \( 1-F(\cdot) \) are the PDF, CDF, and survival function respectively of the transaction process, \( g(\cdot), G(\cdot), \) and \( 1-G(\cdot) \) are the PDF, CDF, and survival function respectively of the latent attrition process, \( n \) is the number of observed repeat transactions, \( T \) is the total observed
time, \( t_j \) indicates the time between each repeat transaction \( j \) from 1 to \( n \), \( t_n \) is the time of the last observed transaction, and \( p \) is the parameter of the geometric attrition process.

**Exponential Distribution**

When PDF \( f(t) \) is assumed to be exponential, \( f(t) = \lambda \exp(-\lambda t) \). Since the exponential distribution has the hazard function \( \lambda \), as discussed in Chapter 2, this PDF is also the hazard function formulation of the exponential distribution. If we combine the exponential PDF with Equation 3.4, we get

\[
L(n, t_n, T|\lambda, p) = [\lambda^n \exp(-\lambda t_n)(1 - p)^n] \times [p + (1 - p)\exp(-\lambda(T - t_n))] .
\]

\( n = 0, 1, 2, \ldots \).

This individual likelihood function of the combined transaction and attrition processes is the same as the MBG/NBD of Batislam, Denizel, and Filiztekin (2007) and the CBG/NBD Hoppe and Wagner (2007).

**Erlang-2 and Erlang-3 Distributions**

Within the CLV marketing modeling literature, Platzer (2008) was the first to propose use of the Erlang-\( k \) distributions for the transaction process. Once we move away from the exponential distribution with its constant hazard rate, things become more interesting. The PDF of the Erlang-2 distribution is \( f(t) = \lambda^2 t \exp(-\lambda t) \), which is equivalent to \( t \sim \text{Gamma}(2, \lambda) \), with CDF \( F(t) = 1 - [\exp(-\lambda t)(1 + \lambda t)] \), and baseline hazard function \( \lambda_0(t) = \frac{\lambda^2 t}{1 + \lambda t} \). The continuous PDF expressed in terms of the hazard function is

\[
f(t) = \frac{\lambda^2 t}{1 + \lambda t} \exp \left( - \int_0^t \frac{\lambda^2 u}{1 + \lambda u} \, du \right) .
\]
Taking the integral in Equation 4.2 gives us the PDF initially described. We need the PDF expressed with the hazard function in order to incorporate time-varying covariates through the hazard function. If we bring Equation 4.2 into Equation 3.4, we get the individual likelihood function

(4.3) \[ L(n, t_n, T|\lambda, p) = \left( \prod_{j=1}^{n} \frac{\lambda^2 (t_j - t_{j-1})}{1 + \lambda (t_j - t_{j-1})} \exp \left( - \int_{t_{j-1}}^{t_j} \frac{\lambda^2 u}{1 + \lambda u} du \right) \right) (1 - p)^{n-1} \times \]

\[ p + (1 - p) \exp \left( - \int_{t_n}^{T} \frac{\lambda^2 u}{1 + \lambda u} du \right), \quad n = 0, 1, 2, \ldots. \]

Once we move past the exponential distribution, we need additional data for the time between each transaction event, including the initial event as the transaction distribution is no longer memoryless.

The application of the Erlang-3 distribution is similar. We have the PDF \( f(t) = \frac{\lambda^3 t^2}{2(1 + \lambda t + (\lambda t)^2)} \exp(-\lambda t) \), which is equivalent to \( t \sim \text{Gamma}(3, \lambda) \), with CDF \( F(t) = 1 - \left\{ \exp(-\lambda t) \left( 1 + \lambda t + (\lambda t)^2 \right) \right\} \), and baseline hazard function \( \lambda_0(t) = \frac{\lambda^3 t}{2(1 + \lambda t + (\lambda t)^2)} \). The continuous PDF of the Erlang-3 distribution, expressed in terms of the hazard function, is

(4.4) \[ f(t) = \frac{\lambda^3 t^2}{2(1 + \lambda t + (\lambda t)^2)} \exp \left( - \int_{0}^{t} \frac{\lambda^3 u}{2(1 + \lambda u + (\lambda u)^2)} du \right). \]

Taking the integral in Equation 4.4 gives us the PDF as expected. Again, we need the PDF expressed with the hazard function in order to incorporate time-varying covariates through the hazard function. If we bring Equation 4.4 into Equation 3.4, we get the individual likelihood function

(4.5) \[ L(n, t_n, T|\lambda, p) = \left( \prod_{j=1}^{n} \frac{\lambda^3 (t_j - t_{j-1})^2}{2 \left( 1 + \lambda (t_j - t_{j-1}) + (\lambda (t_j - t_{j-1}))^2 \right)} \exp \left( - \int_{t_{j-1}}^{t_j} \frac{\lambda^3 u}{2 \left( 1 + \lambda u + (\lambda u)^2 \right)} du \right) \right) (1 - p)^{n-1} \times \]

\[ p + (1 - p) \exp \left( - \int_{t_n}^{T} \frac{\lambda^3 u}{2 \left( 1 + \lambda u + (\lambda u)^2 \right)} du \right), \quad n = 0, 1, 2, \ldots. \]
While Equation 4.5 looks more complex, we really are just plugging in the baseline hazard function for the Erlang-3 distribution, which does have the same time $t$ in multiple different places in its formulation.

For modeling inference via an MCMC sampler, with the exponential, Erlang-2, and Erlang-3 transaction distributions, I assume

$$\lambda \sim \text{Gamma}(r, a),$$

$$p \sim \text{Beta}(s, b),$$

and

$$r, a, s, b$$ are all distributed Gamma with noninformative priors.

The exponential distribution mixed with the Gamma distribution is equivalent to the NBD, at least when there is no ceiling on the number of transactions. The Erlang-2 distribution mixed with the Gamma distribution is equivalent to the CNBD. I note that Platzer’s (2008) CLV model is one of the few stochastic process CLV models that similarly alters the purchases assumptions rather than the lifetime assumptions, which I extend with my CLV modeling framework in the next chapter. Additionally, the CNBD has a fairly long history in stochastic process marketing models (Morrison and Schmittlein 1988). Gupta (1991) describes modeling purchases with a nonstationary Erlang-2 process, though not with a proportional hazards formulation. The Erlang-2 is an important counterpoint to the exponential distribution, because the Erlang distributions are regular in repeat event occurrence whereas the exponential distribution is purely random.

**Individual Likelihood of the Pareto/NBD**

The Pareto/NBD (Schmittlein, Morrison, and Colombo 1987) was the first stochastic process CLV model, making it an important model against which to judge the effectiveness of any new stochastic process CLV model. With the Pareto/NBD we assume exponential distributions for each individual for both the transaction and attrition processes. Fader and Hardie
(2005) provide a derivation of the individual likelihood function, which for convenience I repeat here as Equation 4.6

\[
L(n, t_n, T|\lambda, \mu) = \frac{\lambda^n \mu}{\lambda + \mu} e^{\lambda + \mu t_n} + \frac{\lambda^{n+1} \mu}{\lambda + \mu} e^{(\lambda + \mu T)} , n = 0, 1, 2, ... ,
\]

where \( \mu \) is the rate parameter of the attrition process. For modeling inference of the Pareto/NBD within an MCMC sampler, I assume

\[
\lambda \sim \text{Gamma}(r, \alpha),
\]

\[
\mu \sim \text{Gamma}(s, \beta),
\]

and

\[
r, \alpha, s, \beta \text{ are all distributed Gamma with noninformative priors.}
\]

Fader and Hardie (2005) also provide derivations for \( P(\text{Alive}) \), repeated below as Equation 4.7, as well as for the expected count of future transactions, repeated here as Equation 4.8.

\[
(4.7) \quad P(\text{Alive}) = 1/[1 + \left( \frac{\mu}{\lambda + \mu} \right) \exp(-(\lambda + \mu)(T - t_n)) - 1], \text{ and}
\]

\[
(4.8) \quad \text{Future Transactions} = \frac{\lambda}{\mu} \left( 1 - \exp(\mu t_{\text{fut}}) \right) \times P(\text{Alive}).
\]

Since the underlying individual distributional assumptions of the Pareto/NBD are exponential distributions, we have equations to estimate these amounts with each iteration of an MCMC sampler. With my CLV modeling framework, the estimated counts for future transactions do need to be simulated. I do not have a CLV modeling framework for adding time-varying covariates into the transaction process where the attrition process is a continuous distribution like the Pareto/NBD. In the next chapter, I utilize the Pareto/NBD in the analysis of simulated data in order to have a baseline comparison against which we can compare the more complex models I am proposing to demonstrate my CLV modeling framework.
Description of Empirical Data

For Chapter 5, the empirical data for the modeling applications comes from the IRI Marketing Science dataset (Bronnenberg, Kruger, and Mela 2008) that IRI has made available to academic researchers. The overall dataset contains store sales from 30 categories, 47 markets, with 11 years of data covering 2001 through 2011. The dataset also included consumer panel data from the Eau Claire, WI and Pittsfield, MA markets during the same time period.

Cohort of Soup Customers

Soup is a product category well known for its seasonal demand in the United States, with soup in greater demand during times of colder weather. I chose soup as the product category to examine with my CLV modeling framework because a climate-driven transaction pattern is exogenous in nature. I focus on purchases of soup from one store in the consumer panel, and selected a cohort of every panelist who purchased soup at that store during 2001, for 1,775 households in all. Over the 11 years of data for this cohort, the median amount spent per transaction is $4.08, or 1.405 in log(dollars). For customers with more than one transaction, the correlation between the total number of transactions for each customer and log(dollars) is -0.005 (t = -0.212). In the following empirical studies of soup, I take the viewpoint of store management who would like to make forecasts of expected transactions of their soup customers, forecasts which can be used not just to decide who may be better customers to contact with marketing offers, but potentially for more general management purposes such as category management.
With our cohort of soup purchasers, the observed pattern of transactions is roughly sinusoidal over time (Figure 4.1). While there are various kinds of seasonal purchase patterns dependent on the usage patterns of a product or service, I find the annual cycle of soup purchasing to be a motivating challenge for CLV models. The trend in transactions slightly decreases with time; the linear regression on transactions by week for the 6 years cohort data in Figure 4.1 has a significant slope of -0.316 ($t = -10.44$), indicating that the cohort is on average transacting less over time. Fitting the same data to a Poisson GLM cosine trend model provides parameter estimates of $\beta_1 = 0.437$ ($z = 49.41$), and $\beta_2 = -0.158$ ($z = -18.29$).
Figure 4.2: Number of Soup Customers by Number of Repeat Transactions, 4 Years Observed

Figure 4.3: Average Intertransaction Times by Soup Customers with One or More Repeat Transactions

Figure 4.2 is a snapshot of the number of customers by their number of repeat transactions after observing the cohort for four years. 11.8% of customers have no repeat transactions before January 2005, which by itself is difficult to interpret whether that is an excess
number of zero repeat transactions. The percent of customers with zero repeat transactions decreases from 17.6% after observing the cohort for two years to 8.6% at the end of 11 years, so many of these potential zero-state households in the cohort do transact eventually, if infrequently. Figure 4.3 shows us the distribution of average intertransaction times for customers who do have at least one repeat transaction over the full 11 years of data. The mode is roughly around 2 to 3 months, and only 2.6% of customers have average intertransaction times greater than two years.
CHAPTER 5: THE GENERAL CLV MODELING FRAMEWORK

To date, most of the stochastic process CLV model literature has maintained the original assumption of the Pareto/NBD that a customer's transaction rate remains stationary over time. I have mentioned this before, and it bears repeating, there are many categories of noncontractual goods and services where this stationary purchase rate assumption is violated, particularly categories with seasonal transaction patterns. Additionally, marketing interventions are often nonstationary either in execution or in the way a customer experiences them. For example, various in-store promotions are either present or absent at the times when the customer is visiting a store. As will be demonstrated later in this chapter with simulated data and then with empirical panel data, when customer transaction patterns are nonstationary, current stochastic process CLV models will produce biased estimates of future expected value. This bias is anticipated with existing stochastic process CLV models, as Schmittlein, Morrison, and Colombo (1987, pg. 3) acknowledge, “The [stationary] Poisson assumption would not be as good with catalog purchases since the new editions of the catalog may be mailed out on a very regular schedule.”

In Chapter 3, I highlighted that the generalized form of the individual likelihood function for CLV models that are formulated similarly to the Pareto/NBD with separate transaction and latent attrition stochastic processes, more recently termed latent attrition models, will produce biased estimates when fitted to data that is nonstationary as well as data that has an excess of customers with no repeat transactions. I emphasize this bias is intrinsic to the structure of the individual likelihood function, and not assumptions specific to any one CLV model.

I discuss and illustrate this bias within two sections in this chapter. First, I develop a CLV modeling framework that addresses the two problematic issues in the generalized likelihood function identified in Chapter 3: that of nonstationary transactions and that of zero-repeat customers. I also generalize this framework on its transaction process to not limit it to any one
transaction distribution. Next, I demonstrate how the framework performs with two studies, one study utilizing simulated data, and an empirical study with soup transaction data.

**The General CLV Modeling Framework**

In order to address both the timing bias of estimates from current CLV models due to seasonality as well as the bias introduced by customers with zero repeat transactions, I propose the following three components are necessary in a CLV model:

1) the incorporation of time-varying covariates into the transaction process,
2) a zero-inflated individual likelihood function, and
3) a generalization of the transaction process to allow any probability distribution that has a hazard function which is a function of time.

The first two components of the framework address both the timing bias of estimates from current latent attrition models due to seasonality as well as the parameter bias introduced by customers with zero repeat transactions. The third component incorporates findings from the marketing literature, particularly Platzer and Reutterer (2016), that multiple different distributions are potentially appropriate for modeling the time between transactions.

To avoid the individual integral in the right-censored time period in Equation 3.3, I assume a discrete geometric distribution for the individual attrition process. My framework does not have a closed mathematical form, and solving for the individual integrals required of a continuous distribution for attrition within the right-censored time period would add additional computational challenges. I start the index of the attrition process at 0 (Batislam, Denizel, and Filiztekin 2007) to account for customers who have no repeat transactions. I will demonstrate in the simulation and empirical studies later in this chapter that the 0 index can be improved upon with a zero-inflated individual likelihood specification that addresses any excess of customers.
with no repeat transactions. The initial structure for the individual likelihood of a customer that I extend with my CLV modeling framework is Equation 3.4 above, repeated here for convenience:

\[
L(T, n|f(t), p) = \left[ \prod_{j=1}^{n} f(t_j) \right] (1 - p)^n \times [p + (1 - p)(1 - F(T - t_n))];
\]

\[n = 0, 1, 2, ... ,\]

where \(f(\cdot), F(\cdot),\) and \(1 - F(\cdot)\) are the PDF, CDF, and survival function respectively of the transaction process, \(p\) is the rate parameter of the geometric latent attrition process, \(n\) is the number of observed repeat transactions, \(T\) is the total observed time, \(t_j\) indicates the time between each repeat transaction \(j\) from 1 to \(n\), and \(t_n\) is the time of the last observed transaction.

**Adding Time-Varying Covariates on the Transaction Process Through a Proportional Hazards Formulation**

I described the hazard function above in Chapter 2, and I will be using a proportional hazards formulation of the transaction PDF to add time-varying covariates to my CLV modeling framework. For convenience, I repeat Equations 2.6 and 2.7 below,

\[
f(t) = \lambda(t)[1 - F(t)] = \lambda(t) \exp \left( - \int_{u}^{t} \lambda(u) du \right) \quad \text{and}
\]

\[
\lambda(t) = \lambda_0(t) \exp(x'\beta),
\]

where \(\lambda_0(t)\) is the baseline hazard function \(\frac{f(t)}{1 - F(t)}\). From Lawless (1987), the likelihood of a stochastic process over multiple events becomes:

\[
L(n, t_j, T | \lambda_0(t), \beta) = \left[ \prod_{j=1}^{n} \lambda_0(t_j) \exp(x'\beta) \right] \exp \left( - \int_{0}^{T} \lambda_0(u) \exp(x'\beta) du \right).
\]

In my CLV modeling framework, I maintain the full PDF with the hazard function in the transaction likelihood. Since the hazard function describes the instantaneous rate of an event occurring (Equation 2.2), the proportional hazards method of incorporating covariates allows covariates to modify how the instantaneous rate of the event changes regardless of the distribution
used for the stochastic process. I note that the covariates should not include an intercept term as the mean rate is implicit in the baseline hazard function.

Next, we need to allow for the covariates to vary over time. While $\lambda(t)$ may be continuous, we can discretize the hazard function to the level of granularity in time units, e.g. weeks, available in the empirical data. Within each time unit, the covariates remain invariant, but with enough granularity in time we can approximate a continuous distribution. Over discrete time units, the integral of the hazard function becomes a summation over the time period in question, between 1 and $T$, where $T$ is the end of the observed time period to which the model is fitted:

$\int_0^T \lambda(u) du = \sum_{k=1}^T \lambda_0(t_k) \exp(x_k'\beta)$.  

In the point process literature, the level of granularity used for their proofs is a time unit small enough so that only one event can occur, e.g. Ross (2010). The level of granularity in a latent attrition model with time-varying covariates should be guided by the transactional CDF such that the probability of more than one event in a time unit is uncommon, although there is a tradeoff to be made with what granularity is available in the data to be modeled.

Equation 5.2 can be brought into Equation 5.1 to give us the individual likelihood for a stochastic process with time-varying covariates:

$L(n, t_j, T | \lambda_0(t), \beta) = \prod_{j=1}^n \lambda_0(t_j) \exp(x_j'\beta) \exp[-\sum_{k=1}^T \lambda_0(t_k) \exp(x_k'\beta)]$.

The $j$ and $k$ time indexing in the likelihood requires further explanation. From Equation 5.1, the individual likelihood is a combination of the survival function over the entire observation period with the product of the hazard functions at the time of each event that occurred. There are $n$ repeat transactions between 1 and $T$. The $j$'s are the time of each transaction, so only the covariates active during the time unit of the event are a part of the product term. The $k$'s index each discrete unit time period between 1 and $T$ to allow for the hazards over that time period to be summed together.
Consider modeling the effect of a marketing promotion by adding one covariate. If we assume the promotion only has an effect in one time period, that covariate will be active at a specific time period $k$. However, that effect will carry into later time periods because that one time period will also be a part of the exponentiated summation into later time periods before a transaction occurs. That one time period change to the CDF will be carried on until that next transaction occurs. Now if a transaction occurs in the same time period as the promotion, that transaction will also be active at time period $j$. And with that transaction, the summation is then cleared to start again up to the next transaction.

Bringing Equation 5.3 into Equation 3.4, the individual likelihood for a latent attrition model with time-varying covariates becomes:

\[
L(f(t), p | T, n) = \left[ \prod_{j=1}^{n} \lambda_0(t_j) \exp(x_j \beta) \right] \exp \left[ - \sum_{k=1}^{n} \lambda_0(t_k) \exp(x_k \beta) \right] (1 - p)^n \times \left[ p + (1 - p) \exp \left[ - \sum_{k=t_n+1}^{n} \lambda_0(t_k) \exp(x_k \beta) \right] \right], \ n = 0, 1, 2, \ldots .
\]

Equation 5.4 allows for the inclusion of time-varying covariates to any specification of the transaction process where the distribution has a hazard function as a function of time, i.e. the limit in Equation 2.2 must exist. This formulation of the individual likelihood for a latent attrition model allows for covariates that are not applicable to the customer at time $k$, e.g. the customer is not in the store to see any in-store displays, to not influence the transaction process by setting those covariates to 0 at that time period, but covariates that have been applicable remain a part of the cumulative effect until a transaction happens.

A seasonal correction is needed when the right-censored time period, $T - t_n$, is not i.i.d. across individuals in the data, which I will illustrate in the simulation and empirical studies later this chapter. A seasonal correction will adjust the size of the survival function within the right-censored time period of the individual likelihood function to be less biased by the timing of model fitting. This will directly improve estimates of $P(Alive)$. 

78
Exponentiating the vector $x'\beta$ is common practice in the survival analysis literature for the clarity it provides describing how covariates change the hazard rate. When a $\beta$ parameter is positive, it provides a multiplier effect greater than 1, whereas if the parameter is negative, the covariate reduces the hazard rate with a multiplicative effect between 0 and 1. Hazard rates can be decreasing over time but never negative, and the floor of 0 is necessary. The tradeoff to gain that clarity is that the summation of exponentials is computationally taxing. I note that when a covariate is a simple dummy variable active during a time unit $j$ or $k$, which is a reasonable parameterization for many kinds of seasonal corrections, we can substitute $\gamma = \exp(1'\beta)$ into the likelihood to reduce the computational complexity.

Zero-Inflated Individual Likelihood

The zero-inflated model was first proposed in the marketing literature by Morrison (1969). When we have count data with more zeroes than would be expected from a given count distribution, the zero-inflated model incorporates an additional probability of a zero count to the model likelihood to account for the excess zeroes in the observed data. Morrison suggested with simulation studies that without the zero-inflated correction, light transactions will be underestimated and heavy transactions will be overestimated because the inferred model parameters would be biased. The structure of the zero-inflated model for the probability of observing $n$ counts is

$$p(n) = \begin{cases} P(n = 0) = \omega + (1 - \omega)f(0) \\ P(n > 0) = (1 - \omega)f(n) \end{cases}, \quad 0 < \omega < 1,$$

where $f(\cdot)$ is the distribution for 0 or more than 0 observed counts. The concept behind the formulation of the zero-inflated specification is to model a zero-state that allows for what Morrison terms a hardcore group of nonusers. In the context of a CLV or latent attrition model,
the zero-inflated specification allows us to estimate future transactions from customer data that includes a sizeable number of customers with no observed repeat transactions.

The labeling of the zero-state as a group of hardcore nonusers is perhaps too strongly worded. Behaviorally, the zero-state is different from the attrition process, and combining Equation 5.5 with Equations 3.3 or 3.4 isolates the zero-state from the attrition process. It is possible that customers with no repeat transactions are new to the product category and there is extensive literature that initial product trial behavior is different from repeat usage behavior (e.g., Fader, Hardie, and Huang 2004). It is also possible that something random happened to influence a singular household purchase with no further observed transactions. The classic example is the visiting mother-in-law. Panel data does not often include details regarding recruitment. In either case, the attrition process would not correctly represent many customers with no repeat transactions. Instead, I focus on the fact that individuals with zero repeat transactions by definition have no observed time between transactions. Thus, they contribute no information to the first half of the individual likelihood function of Equations 3.3 or 3.4. By specifying what is effectively a separate latent state in the individual likelihood for customers with no observed repeat transactions, we can focus inference of the model parameters more on customers who do have repeat transactions. This separation reduces the overall bias of parameter estimates. I demonstrate in the simulation and empirical studies that the addition of a zero-inflated specification to a CLV model greatly improves aggregate estimates of \( P(Alive) \) and sometimes estimates of future transactions as well.

Extending Equation 5.4 with the zero-inflated specification gives us the following for the individual likelihood that includes time-varying covariates as well as a zero-inflated correction.

\[
\begin{align*}
P(n = 0) &= \omega + (1 - \omega) \left[ p + (1 - p) \exp\left[ - \sum_{k=1}^{T} \lambda_0(t_k) \exp(x_k'\beta) \right] \right], \text{ and} \\
P(n > 0) &= (1 - \omega) \left[ \prod_{j=1}^{n} \lambda_0(t_j) \exp(x_j'\beta) \right] \exp\left[- \sum_{k=1}^{n} \lambda_0(t_k) \exp(x_k'\beta) \right] (1 - p)^n \times \\
&\quad \left[ p + (1 - p) \exp\left[ - \sum_{k=1}^{n+1} \lambda_0(t_k) \exp(x_k'\beta) \right] \right].
\end{align*}
\]
where \( n = 0, 1, 2, \ldots; 0 < \omega < 1 \). With a latent attrition model, \( n \) counts the number of repeat transactions by an individual. That is, the number of transactions after their first transaction. The first transaction gives us the time of entry in the dataset for an individual, but does not by itself provide us with the time between transactions that we need to properly calibrate the model. If a customer has zero observed repeat transactions, then the probability the customer is in the zero state is described as (Rodrigues 2003)

\[
P(n = 0, \omega) = \frac{\omega}{\omega + (1-\omega)f(0)}, 0 \leq \omega < 1.
\]

I term this probability \( P\text{(Zero-State)} \).

In my MCMC implementation, I make the assumption that customers in the zero-state that iteration have \( P\text{(Alive)} = 0 \). When the time between the only instance the customer is observed and the end of the observation period is longer than for most customers and longer than a seasonal cycle or two, it is highly unlikely that customer would continue with a sizeable number of transactions in the future. Should that customer have their first repeat transaction at a later date, which is not unusual in empirical data, the amount they will go on to transact is usually extremely limited. My implementation of the zero-inflated specification within the context of MCMC inference is described in Appendix A.

As discussed in Chapter 3, we are working with the right-censored portion of individual likelihood function: \[ p + (1-p)(1 - F(T - t_n)) \]. With enough time in the observed data, \( F(T - t_n) \) will approach 1, such that only \( p \) is material. In other words, I assume that with enough observed time, the customers with no observed repeat transactions should be treated as inactive. With a zero-inflated specification, the proportion of customers that have no repeat transactions is thus approximately \( \omega + (1-\omega)E[p] \). Otherwise, the proportion of zero-repeat customers per the model without the zero-inflated specification is approximately \( E[p] \). These are approximations to help the researcher to determine the need for the zero-inflation correction. If a dataset has an excess of customers with no repeat transactions, an analyst can estimate a latent
attrition CLV model without the zero-inflation specification. If $E[p]$ is less than the actual proportion of customers with zero repeat transactions, that is indicative there is an excessive number of customers with no repeat transactions in the data.

The three components of the general CLV modeling framework — correcting nonstationary (i.e. seasonal) transaction data through time-varying covariates, a zero-inflation specification, and generalization of the transaction distribution, are incorporated into the individual level likelihood function with Equation 5.6. Regardless of the analyst’s choice of the prior or mixing distributions to model population heterogeneity, or the method of statistical inference utilized, these three components of the framework are necessary to reduce the overall bias of CLV model estimates of future transactions as well as the probability of whether a customer is still active. In the next section, I make additional modeling assumptions specific to the simulation and empirical studies to demonstrate the framework, but I emphasize the generality of the components of the CLV modeling framework.

**Simulation and Empirical Studies**

In this section, I apply the CLV modeling framework I developed above first to simulated data, and then to empirical panel data. The simulated data is fitted with the different models discussed in the previous chapter, i.e. the Pareto/NBD, and then the framework models with the exponential, Erlang-2, and Erlang-3 transaction distributions. This simulation study highlights the benefits as well as some of the remaining challenges with CLV models. With the empirical study that follows the simulation study, I focus on the exponential and Erlang-2 transaction distributions. The simulation study shows the Pareto/NBD does not perform as well as the other models. Additionally, I had some difficulty getting the Erlang-3 models to converge with the empirical transaction data, leaving nothing for comparison.
I start by extending Equation 5.6 with the exponential and Erlang-2 transactions distributions. The baseline hazard function of the exponential distribution is the constant $\lambda$, which by Equation 5.6 gives us

\begin{equation}
P(n = 0) = \omega + (1 - \omega) \left[ p + (1 - p) \exp \left\{ - \sum_{k=1}^{T} \lambda \exp(x_k' \beta) \right\} \right], \text{ and}
\end{equation}

\begin{equation}
P(n > 0) = (1 - \omega) \left[ \lambda^n \prod_{j=1}^{n} \exp(x_j' \beta) \right] \exp \left\{ - \sum_{k=1}^{T} \lambda \exp(x_k' \beta) \right\} (1 - p)^n \times \left[ p + (1 - p) \exp \left\{ - \sum_{k=t_n+1}^{T} \lambda \exp(x_k' \beta) \right\} \right],
\end{equation}

where $n = 0, 1, 2, \ldots; 0 < \omega < 1; \lambda > 0$.

The Erlang-2 distribution was originally proposed for latent attrition models by Platzer (2008) to fit transaction data where the time between transactions was not distributed exponentially. Platzer and Reutterer (2016) have more recently demonstrated use of the two-parameter Gamma distribution with the shape parameter inferred for each customer. I discuss the Erlang-2 in more depth here to demonstrate how the time indices in the individual likelihood are affected when the baseline hazard function is not a constant over time.

The PDF of the Erlang-2 distribution is $f(t) = \lambda^2t \exp(-\lambda t)$. The continuous PDF expressed in terms of the hazard function is was previously described in the last chapter in Equation 4.2. We need the PDF expressed with the hazard function in order to incorporate time-varying covariates. If we bring the Erlang-2 hazard function into Equation 5.6, we get the individual likelihood

\begin{equation}
P(n = 0) = \omega + (1 - \omega) \left[ p + (1 - p) \exp \left\{ - \sum_{k=1}^{T} \frac{\lambda^2 t_k}{1 + \lambda t_k} \exp(x_k' \beta) \right\} \right], \text{ and}
\end{equation}

\begin{equation}
P(n > 0) = (1 - \omega) \left[ \prod_{j=1}^{n} \frac{\lambda^2(t_j - t_{j-1})}{1 + \lambda(t_j - t_{j-1})} \exp(x_j' \beta) \right] \exp \left\{ - \sum_{k=1}^{T} \sum_{j=1}^{t_k} \frac{\lambda^2(t_k - t_{j-1})}{1 + \lambda(t_k - t_{j-1})} \exp(x_k' \beta) \right\} (1 - p)^n \times \left[ p + (1 - p) \exp \left\{ - \sum_{k=t_n+1}^{T} \frac{\lambda^2(t_k - t_{n})}{1 + \lambda(t_k - t_{n})} \exp(x_k' \beta) \right\} \right],
\end{equation}

where $n = 0, 1, 2, \ldots; 0 < \omega < 1; \lambda > 0$. 

83
I introduce here a double summation into the individual likelihood to account for the time between repeat transactions, $t_k - t_{j-1}$. Without covariates, the time between transactions remains stationary. In order to maintain the stationary assumption of the transaction process, time $t$ needs to be reset to 1 (not 0 since we are working with discrete time windows rather than continuous time) after each transaction occurs. Effectively, we have a single, outer, summation of the time between 1 and $t_n$, the time of the last observed transaction. But we need to reset time $t$ after each transaction. That is what the second, inner, summation represents. After each event, we count from 1 until the next event happens and, over $n$ events, we sum up that over all of the observed past. This indexing only affects time $t$ within the hazard function as time has passed since the last transaction, and not the indexing of the time-varying covariates. With the time-indexing introduced with the Erlang-2, extending this discussion to the Erlang-3 is left as an exercise for the reader (as the euphemism goes).

Population Heterogeneity and Inference

In Equations 5.8 and 5.9, $\lambda$ is an individual level parameter for the transaction process, $p$ is an individual level parameter for the attrition process, $\omega$ is a population level parameter for the zero-inflated specification, and $\beta$ is a vector of population level covariate parameters. To incorporate population heterogeneity for the transaction and attrition processes, I extend Equations 5.8 and 5.9 as hierarchical Bayesian models, I assume the following:

$\lambda \sim \text{Gamma}(r, a),$

$p \sim \text{Beta}(s, b),$

$r, a, s, b$ are separately distributed Gamma with noninformative priors,

$\omega$ is distributed Beta with noninformative priors, and

$\beta$ is distributed with a noninformative, uniform prior of 1.
I utilize a Gibbs sampler that is primarily Metropolis within Gibbs sampling (Rossi, Allenby, and McCulloch 2005) for most parameter estimates. The sampler is described in detail in Appendix A. Since there is not a closed form solution for a count of future transactions, I simulate future transactions to produce an expectation at an individual level. The CDF-based simulation to generate future transactions is described in Appendix A.

**Study Design**

To evaluate how well the general CLV modeling framework addresses seasonal timing bias as well as the bias from customers with no repeat transactions, for the simulation study I utilize a (2 x 2 x 3 + 1) x 4 design with a seasonal correction (yes / no), a zero-inflation specification (yes / no), three different transactional distributions (exponential / Erlang-2 / Erlang-3), and then the Pareto/NBD for thirteen different model conditions in all. These 13 models are then fitted to simulated data that uses a seasonal covariate of one of four different magnitudes. The model with no seasonal correction, no zero-inflation, and the exponential transaction distribution is equivalent to the MBG/NBD (Batislam, Denizel, and Filiztekin 2007).

To assess the accuracy of the estimates of future transactions and $P(Alive)$, I repeatedly fit the models over a two-year time period every eight weeks for 14 separate model fitting times. For the simulation studies, this requires 182 model fittings for the 13 different model conditions. With the empirical soup transaction data, the observed data for fitting the model begins with the start of the consumer panel in January 2001 up to the time of model fitting. Figure 5.1 shows graphically the soup transaction data used for model fitting as well as the holdout data that will be used as a comparison to the model estimates. The simulation study follows the same design.
The point of the repeated model fittings is to make salient the bias that results from not using the elements of the CLV modeling framework. I generate estimates of \( P(\text{Alive}) \) at the time of model fitting as well as estimate the total number of future transactions by the cohort from the time of model fitting through the following 104 weeks. Looking forward two years is a longer time window for estimates than most published latent attrition models.

The models are estimated primarily in C, with the data sent to the models through R’s \( \text{Call} \) interface.\(^3\) It is generally much more complex to incorporate time-varying covariates than time-invariant covariates into any time-to-event model, and there is a computational tradeoff for the improved model estimates. I note that seasonal corrections that can be added through simple dummy variables, which is how the simulation study is designed, require less computation time.

With the individual likelihood functions above, and for the Pareto/NBD individual likelihood detailed in Chapter 4, we can estimate the posterior distribution for \( P(\text{Alive}) \) at the end of the observation period by means of Equations 3.7 and 4.7 respectively. Estimating future

\(^3\) Because of the sum of exponentials to incorporate time-varying covariates, as well as the use of simulation to estimate future transactions, the general CLV modeling framework is computationally intensive, and I strongly recommend the use of a compiled computer language for model inference.
transactions is another story. With the Pareto/NBD, we do have Equation 4.8 to calculate future transactions for each individual. With the CLV modeling framework, we do not have a closed form to estimate future transactions. For example, there is no count distribution for a nonstationary log-logistic distribution that I know of. Instead, I use a CDF-based simulation to calculate this for each individual with each iteration of the MCMC chain, as detailed in Appendix A. With enough granularity, this is not an issue. But if the discrete time periods utilized are not granular enough, then we will underestimate the number of future transactions. This issue does come up in the simulation study.

Simulation Study Results

I examine my CLV modeling framework first with a simulation study in order to explore parameter recovery when the three elements of my CLV modeling framework – correction for seasonality through time-varying covariates, zero-inflation specification, and specification of the transaction distribution – are included within a CLV model. The simulation study will also demonstrate improved estimates of simulated customer future transactions and whether they remain active. The study is designed to illustrate the biased estimates from CLV models that do not correct for nonstationary transaction patterns or customers with no repeat transactions.

I emphasize the word “improved.” The CLV modeling framework does improve upon existing CLV models, but we are not hitting all of the true values of the parameter values and estimates spot-on. There remain statistical inference challenges related to the joint inference of both the transaction and attrition processes over the right-censored time portion of the individual likelihood. This is particularly true with transaction distributions like the Weibull that I discuss in the next chapter. I believe this part of the likelihood functions affects our ability to infer the underlying true values. That is a future research challenge, but for now the simulation study can be used to show both the improvements as well as the ongoing inference challenges. The
simulation study also highlights some granularity issues for the transaction process, even with a
time unit of one week.

For each of the simulated datasets, I simulate the transactions by means of the transaction
CDF and possible attrition of 1,000 individuals over a six-year period, where each year has 52
weeks, and one week is the smallest time unit in the simulation. Each individual follows an
Erlang-2 transaction process and an independent geometric attrition process. Since this is a
simulation, the attrition process is not latent to me. The last 8 weeks of each year are modified by
a dummy covariate with a population parameter for the strength of the effect.

The individual transaction and attrition rates are drawn from the following distributions
to represent population heterogeneity:

\[
\lambda \sim \text{Gamma}(4, 64), \text{ such that } E[\lambda] = 1/16, \text{ or } 0.0625, \text{ with } SD[\lambda] = 0.03125, \text{ and }
\]

\[
p \sim \text{Beta}(8, 160), \text{ such that } E[p] = 1/21, \text{ or } \sim 0.0476, \text{ with } SD[p] \sim 0.0164.
\]

The base value of the \( \lambda \)'s represents the exponential distribution. I multiply the drawn \( \lambda \)'s by 2 to
represent the Erlang-2 parameters. This will also make it easier to compare the inferred values
from the different transaction distributions used in the modeling process.

Note the implied values of the standard deviations here based on the distribution
parameters used. I have had reasonable results coming close to the implied mean of the
transaction and attrition population distributions with simulated data that has smaller variance for
the individual parameters. Recovering the implied variance of the population mixing
distributions remains an ongoing challenge. With empirical data, the overall range of the
individual rates is potentially fairly wide. I wanted to demonstrate with this simulation study that
CLV models are reasonably robust to population heterogeneity with mean values, if not exact.
Giving these models a challenge will highlight where there remain some issues.

The simulation seasonal covariate, using the substitution \( \gamma = \exp(1'\beta) \) described above,
is 1 of 4 values, either 0.5, 1, 1.5, and 3, where \( \gamma = 1 \) is equivalent to no seasonal effect. For each
seasonal covariate, I simulate 25 datasets to model, for 100 datasets in all. Time 0 represents time of entry, so all simulated transactions represent repeat transactions. This gives us 25 datasets with a depressive seasonal effect, 25 with no seasonal effect, and two sets of 25 datasets where the seasonal effect increases transactions. There is a chance of attrition with each transaction, so the datasets with the depressive seasonal effect will see less attrition than the datasets with an increased seasonal effect. Finally, to demonstrate that an excess number of zero-repeat transaction customers biases the estimates of existing models, there is a 15% chance that an individual will leave before they even transact. Figure 5.2 shows an example of simulated weekly transactions where the seasonal parameter equals 3.

*Figure 5.2: Simulated Transactions Example, Erlang-2 Transaction Process, γ = 3*

To illustrate how my CLV modeling framework corrects for the bias of existing CLV models, I follow Figure 5.1 where I refit the models every 8 weeks from between two to four years for each dataset, estimating \( P(\text{Alive}) \) at each model fitting and forecasting future transactions for the next 104 weeks. The model with the Erlang-2 transaction distribution, zero-
inflation specification, and a dummy variable to represent seasonality, that matches the simulation design, is expected to have the most stable parameter and individual behavior estimates over the simulation.

I only run one MCMC chain for each model. For reduced computational complexity, I use the true population values as the starting values of the MCMC chain. Each chain is run for 5,000 iterations, with the first 2,000 discarded as burn-in. I keep this relatively small as there are 13 models and 4 separate seasonal conditions in all to examine. We have the Pareto/NBD, and then for the models with the geometric attrition process we have three transaction distributions (exponential, Erlang-2, and Erlang-3), zero-inflation specification (yes / no), and correction for seasonality (yes / no). For each set of parameters used for the simulation, there are 4,550 model fittings in all (13 models * 14 time periods * 25 datasets). While the average model run is approximately 18 seconds, that is 81,900 seconds for 25 datasets, or ~ 23 hours. This is repeated four times for each of the four seasonal parameter values simulated. I examine how well the MCMC chains converge in Appendix C. While these limitations, like shortened MCMC chains or the discrete time granularity of weeks, are not ideal, given the computational constraints this simulation study still provides an overall view of how the different models perform.

I focus the model results at the level of the customer base rather than at the level of the individual customer. To compare model estimates of $P(Alive)$ to the customer base, I use the actual proportion of customers active at the time of model fitting. I assume a customer is active if they had at least one transaction in the data during the 104-week time period after model fitting. With the simulation, actual attrition can be recorded, but we need to be cognizant of what can be observed by managers. For both future transactions and $P(Alive)$, I compare model estimates with the actual aggregate values. The aggregate results are consistent with estimating the components of customer equity (Gupta, Lehmann, and Stuart 2004).

For the simulation study, I compare model performance with two summary statistics, the mean bias of the estimate and the root mean squared error (RMSE), to compare model estimates
to the actual aggregate amounts. Each of these statistics are compared by their means over the 25 simulated datasets, looking at each of the 14 time periods to see how the estimates change over time. These statistics are plotted over time with confidence intervals for the 25 simulated datasets with each seasonal covariate. With the simulation study results, most of these plots are in color to help make the different models stand out more clearly. I also provide tables of the overall mean RMSE for each model by seasonal covariate.

I start with the recovery of the proportion of customers who have no observed repeat transactions. The probability underlying the simulation is 0.15, but that is just a probability, and will vary with each dataset. With only 25 simulated datasets for each seasonal covariate, the overall simulation mean will not be exactly 15%. Add to that the underlying variance of the transaction rate parameter over the 1,000 individuals, and some individuals who have not left may not see a repeat transaction during the simulated six years. However, with a larger seasonal covariate parameter, more transactions occur. So we should see fewer individuals with no repeat transactions over time. This is the pattern we see in Figure 5.3. The larger the seasonal parameter, the lower the proportion of individuals with no repeat transaction. And holding the seasonal pattern steady, the proportion drops over time as more individuals have an opportunity to transact.
The inference of the attrition rate parameter affects our estimates of whether an individual remains active, so I start by looking at how well we can recover the proportion of individuals who do not yet have a repeat transaction at the end of each observation period. As stated above, with a zero-inflated specification, the proportion of customers that have no repeat transactions is approximately $\omega + (1-\omega)E[p]$. Otherwise, the proportion of zero-repeat customers is approximately $E[p]$. To check recovery, I plot the mean bias, with its 95% confidence interval, as well as the RMSE, over each of the 14 time periods by model in Figures 5.4 through 5.6, with the mean bias in Figures a and the RMSE in Figures b, that respectively show the exponential, Erlang-2, and Erlang-3 transaction distributions. The Pareto/NBD is not included since it does not have an explicit model for customers with no repeat transactions.
Figure 5.4a: Recovery of Zero-Repeat Customers, Exponential Transactions (Bias)
Figure 5.4b: Recovery of Zero-Repeat Customers, Exponential Transactions (RMSE)
Figure 5.5a: Recovery of Zero-Repeat Customers, Erlang-2 Transactions (Bias)
Figure 5.5b: Recovery of Zero-Repeat Customers, Erlang-2 Transactions (RMSE)
Figure 5.6a: Recovery of Zero-Repeat Customers, Erlang-3 Transactions (Bias)
Figure 5.6b: Recovery of Zero-Repeat Customers, Erlang-3 Transactions (RMSE)
Table 5.1: Recovery of Zero-Repeat Customers, RMSE

<table>
<thead>
<tr>
<th>Transaction Distribution</th>
<th>Parameter by Model</th>
<th>3</th>
<th>1.5</th>
<th>1</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exponential</strong></td>
<td>Base</td>
<td>0.068222 (_{z,s}^b)</td>
<td>0.068901 (_{z,s}^b)</td>
<td>0.070368 (_{z,s}^b)</td>
<td>0.071841 (_{z,s}^b)</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>(0.000439^b_x)</td>
<td>(0.000534^b_x)</td>
<td>0.000604 (_{z,s}^b)</td>
<td>(0.000687^b_x)</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>0.065960 (_{z,s}^b)</td>
<td>0.067746 (_{z,s}^b)</td>
<td>0.070687 (_{z,s}^b)</td>
<td>0.071269 (_{z,s}^b)</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>0.000465 (_{z,s}^b)</td>
<td>0.000563 (_{z,s}^b)</td>
<td>(0.000582^b_x)</td>
<td>0.000724 (_{z,s}^b)</td>
</tr>
<tr>
<td><strong>Erlang-2</strong></td>
<td>Base</td>
<td>0.067309 (_{z,s}^b)</td>
<td>0.066513 (_{z,s}^b)</td>
<td>0.067648 (_{z,s}^b)</td>
<td>0.068583 (_{z,s}^b)</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>0.000462 (_{z,s}^b)</td>
<td>0.000457 (_{z,s}^b)</td>
<td>(0.000447^b_x)</td>
<td>0.000475 (_{z,s}^b)</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>0.066034 (_{z,s}^b)</td>
<td>0.066576 (_{z,s}^b)</td>
<td>0.067850 (_{z,s}^b)</td>
<td>0.068881 (_{z,s}^b)</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>(0.000460^b_x)</td>
<td>(0.000455^b_x)</td>
<td>0.000460 (_{z,s}^b)</td>
<td>(0.000460^b_x)</td>
</tr>
<tr>
<td><strong>Erlang-3</strong></td>
<td>Base</td>
<td>0.066670 (_{z,s}^b)</td>
<td>0.066266 (_{z,s}^b)</td>
<td>0.067751 (_{z,s}^b)</td>
<td>0.068143 (_{z,s}^b)</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>(0.000495^b_x)</td>
<td>(0.000494^b_x)</td>
<td>0.000493 (_{z,s}^b)</td>
<td>(0.000484^b_x)</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>0.066123 (_{z,s}^b)</td>
<td>0.066517 (_{z,s}^b)</td>
<td>0.067571 (_{z,s}^b)</td>
<td>0.068383 (_{z,s}^b)</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>0.000511 (_{z,s}^b)</td>
<td>0.000511 (_{z,s}^b)</td>
<td>(0.000484^b_x)</td>
<td>0.000509 (_{z,s}^b)</td>
</tr>
</tbody>
</table>

\(b, z, s, sz\) indicate statistical significance (p < 0.05) compared to the base, zero-inflated, seasonal, and seasonal/zero-inflated models respectively while holding the transaction distribution and seasonal parameter equivalent.

As can be seen, the models with a zero-inflation model specification have almost no bias and no error. Table 5.1 shows just how small the mean RMSE over the 14 time periods is for the zero-inflated models. Recovering the true proportion of zero-repeat individuals is particularly important because within the MCMC sampler, as detailed in Appendix A, I am using a new method to achieve convergence with a zero-inflated model specification. Individuals with no repeat transactions have no observed data on which to infer their individual transaction and attrition rate parameters. To work around this lack of individual data, I use the mean rates from the last MCMC iteration in order to compute a probability as to whether an individual with no repeat transactions is in the zero-state. The individual difference is the length of the right-censored time period, which does vary in the empirical study, if not the simulation. By using expected values with the mean rates, this method is similar to the ECM (Expectation – Conditional Maximization) algorithm described by Gelman et al. (2014), though applied within a Gibbs sampler.
As predicted, without the zero-inflation specification, the models will underestimate the proportion of customers who have no repeat transactions, and that biases the inference of the attrition process. To examine this bias, I look at the estimated mean of the attrition rate, $E[p]$, which again does not apply to the Pareto/NBD. Figures 5.7 through 5.9 that respectively show the estimated mean bias (Figures a), with the 95% confidence intervals, and RMSE (Figures b) for $E[p]$ as calculated for each dataset (not the bias against the value of $1/21$ underlying the simulation) for the exponential, Erlang-2, and Erlang-3 transaction distributions. The Erlang-2 and Erlang-3 models with the zero-inflation specification come close to the true values here, while the models without the zero-inflation specification overestimate $E[p]$. Now the value of $E[p]$ going into the simulation is $\sim 0.0476$. Overestimating this amount by 0.03 – 0.04 is a sizeable, as well as significant, overestimate. Table 5.2 shows the mean RMSE over the 14 time periods, and confirms that the error of the models without the zero-inflation specification is much larger.
Figure 5.7a: Recovery of $E[p]$, Exponential Transactions (Bias)
Figure 5.7b: Recovery of $E[p]$, Exponential Transactions (RMSE)
Figure 5.8a: Recovery of $E[p]$, Erlang-2 Transactions (Bias)
Figure 5.8b: Recovery of $E[p]$, Erlang-2 Transactions (RMSE)
Figure 5.9a: Recovery of $E[p]$, Erlang-3 Transactions (Bias)
Figure 5.9b: Recovery of $E[p]$, Erlang-3 Transactions (RMSE)
Table 5.2: Recovery of $E[p]$, RMSE

<table>
<thead>
<tr>
<th>Transaction Distribution</th>
<th>Seasonal Parameter by Model</th>
<th>3</th>
<th>1.5</th>
<th>1</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>Base</td>
<td>0.03198 $^{b,z,s,z}$</td>
<td>0.03363 $^{z,s,z}$</td>
<td>0.03613 $^{z,s,z}$</td>
<td>0.03761 $^{z,s,z}$</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>0.00902 $^{b,s}$</td>
<td>0.00917 $^{b,s}$</td>
<td><strong>0.00836</strong> $^{b,s}$</td>
<td><strong>0.00813</strong> $^{b,s}$</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>0.03402 $^{b,z,s,z}$</td>
<td>0.03479 $^{z,s,z}$</td>
<td>0.03563 $^{z,s,z}$</td>
<td>0.03833 $^{z,s,z}$</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td><strong>0.00706</strong> $^{b,s}$</td>
<td><strong>0.00848</strong> $^{b,s}$</td>
<td>0.00850 $^{b,s}$</td>
<td>0.00919 $^{b,s}$</td>
</tr>
<tr>
<td>Erlang-2</td>
<td>Base</td>
<td>0.03261 $^{z,s,z}$</td>
<td>0.03591 $^{z,s,z}$</td>
<td>0.03849 $^{z,s,z}$</td>
<td>0.04047 $^{z,s,z}$</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>0.00460 $^{b,s}$</td>
<td>0.00369 $^{b,s}$</td>
<td><strong>0.00261</strong> $^{b,s}$</td>
<td>0.00377 $^{b,s}$</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>0.03383 $^{z,s,z}$</td>
<td>0.03570 $^{z,s,z}$</td>
<td>0.03830 $^{z,s,z}$</td>
<td>0.04041 $^{z,s,z}$</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td><strong>0.00328</strong> $^{b,s}$</td>
<td><strong>0.00334</strong> $^{b,s}$</td>
<td>0.00267 $^{b,s}$</td>
<td><strong>0.00347</strong> $^{b,s}$</td>
</tr>
<tr>
<td>Erlang-3</td>
<td>Base</td>
<td>0.03304 $^{z,s,z}$</td>
<td>0.03599 $^{z,s,z}$</td>
<td>0.03818 $^{z,s,z}$</td>
<td>0.04092 $^{z,s,z}$</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>0.00422 $^{b,z,s,z}$</td>
<td>0.00343 $^{b,z,s,z}$</td>
<td>0.00362 $^{b,z,s,z}$</td>
<td>0.00552 $^{b,z,s,z}$</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>0.03362 $^{z,s,z}$</td>
<td>0.03566 $^{z,s,z}$</td>
<td>0.03837 $^{z,s,z}$</td>
<td>0.04069 $^{z,s,z}$</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td><strong>0.00303</strong> $^{b,z,s}$</td>
<td><strong>0.00335</strong> $^{b,s}$</td>
<td><strong>0.00338</strong> $^{b,s}$</td>
<td><strong>0.00458</strong> $^{b,s}$</td>
</tr>
</tbody>
</table>

$b, z, s, sz$ indicate statistical significance ($p < 0.05$) compared to the base, zero-inflated, seasonal, and seasonal/zero-inflated models respectively while holding the transaction distribution and seasonal parameter equivalent.

These are the mean values, and the models with a zero-inflated specification come very close to the true values. The simulation mean implied by the true model parameters is 0.04762. The expected value from all of the Erlang-2 model runs where the seasonal parameter equals 3, corrected for seasonality and zero-inflation, is 0.04656, an underestimate of about 2.2%, which is reasonably close. Variance is another story. The simulation variance implied by the true model parameters is 0.0002684. The expected variance from all of the same Erlang-2 model runs corrected for seasonality and zero-inflation is 0.0003642, an overestimate of 36%. I highlight this as one example; in general, my experience has been that these models tend to overestimate the variance with simulated data. There are only two CLV simulation studies I am aware of that reports the variance estimated from simulated data. The first is Korkmaz, Fok, and Kuik (2014), who in one of their two simulated latent classes report the variance of the attrition rate has nearly doubled. The second is Platzer and Reutterer (2016) who only report estimated parameters for five of their 160 simulation scenarios. In the five simulations they report, four of their studies report both lower means (as much as 43% lower) and lower variance that is one case is 73% less.
In other words, the Pareto/GGG of Platzer and Reutterer generally recovers a lower attrition rate, implying longer lifetimes than the true values, with much less variance on those estimates. This is not to single out those who have reported simulation studies, and is an issue that needs to be addressed in future research.

Figure 5.10: Mean Proportion of Customers Still Active

Biased inference of the mean values by uncorrected CLV models goes on to affect the estimates of $P(Alive)$. The proportion of individuals active decreases over time. By model assumption (and as coded into the simulation), individuals have a probability of leaving after each transaction. With seasonal time periods that have increased transaction activity, we will see more individuals leaving over the same period of time. This is shown in Figure 5.10, particularly when
the seasonal parameter equals 3 and $P(Alive)$ declines more quickly during the season the covariate is active. Figures 5.11 through 5.13 respectively show the estimated mean bias (Figures a), with the 95% confidence intervals, and RMSE (Figures b) for $P(Alive)$ as calculated for each dataset for the exponential, Erlang-2, and Erlang-3 transaction distributions.

*Figure 5.11a: Recovery of $E[P(Alive)]$, Exponential Transactions (Bias)*
Figure 5.11b: Recovery of $E[P(\text{Alive})]$, Exponential Transactions (RMSE)
Figure 5.12a: Recovery of $E[P(Alive)]$, Erlang-2 Transactions (Bias)
Figure 5.12b: Recovery of $E[P(\text{Alive})]$, Erlang-2 Transactions (RMSE)
Figure 5.13a: Recovery of $E[P(\text{Alive})]$, Erlang-3 Transactions
Figure 5.13b: Recovery of $E[P(\text{Alive})]$, Erlang-3 Transactions (RMSE)
Table 5.3: Recovery of $E[P(\text{Alive})]$, RMSE

<table>
<thead>
<tr>
<th>Transaction Distribution</th>
<th>Seasonal Parameter by Model</th>
<th>3</th>
<th>1.5</th>
<th>1</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>Base</td>
<td>0.04090 $^b$</td>
<td>0.03782 $^{b,c,z}$</td>
<td>0.03155 $^{c,z}$</td>
<td><strong>0.02809</strong> $^{b,c,z}$</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>0.05375 $^b$</td>
<td>0.05412 $^{b,z}$</td>
<td>0.04933 $^{b,z}$</td>
<td>0.04703 $^{b,z}$</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td><strong>0.03077</strong> $^{b,c,z}$</td>
<td><strong>0.03353</strong> $^{b,c,z}$</td>
<td><strong>0.03152</strong> $^c$</td>
<td>0.03254 $^{b,z}$</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>0.04126 $^c$</td>
<td>0.04947 $^{b,z}$</td>
<td>0.04950 $^b$</td>
<td>0.05309 $^b$</td>
</tr>
<tr>
<td>Erlang-2</td>
<td>Base</td>
<td>0.02094 $^{c,z}$</td>
<td>0.01250 $^{c,z}$</td>
<td>0.01360 $^{c,z}$</td>
<td>0.02005 $^{c,z}$</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>0.02201 $^{c,z}$</td>
<td>0.01462</td>
<td><strong>0.01042</strong> $^b$</td>
<td>0.01514</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>0.01123 $^{b,z}$</td>
<td><strong>0.01210</strong> $^{b,z}$</td>
<td>0.01337</td>
<td>0.01467</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td><strong>0.01048</strong> $^{b,z}$</td>
<td>0.01264</td>
<td>0.01045 $^b$</td>
<td><strong>0.01194</strong> $^b$</td>
</tr>
<tr>
<td>Erlang-3</td>
<td>Base</td>
<td>0.02529 $^{c,z}$</td>
<td>0.02178 $^{c,z}$</td>
<td>0.02752 $^{c,z}$</td>
<td>0.03469 $^{c,z}$</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>0.02325 $^{c,z}$</td>
<td><strong>0.01632</strong> $^{b,z}$</td>
<td>0.01996 $^{b,z}$</td>
<td>0.02645</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>0.01888 $^{c,z}$</td>
<td>0.02357 $^{c,z}$</td>
<td>0.02686 $^{c,z}$</td>
<td>0.02895 $^{c,z}$</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td><strong>0.01393</strong> $^{b,z,c}$</td>
<td>0.01726 $^{b,z}$</td>
<td><strong>0.01905</strong> $^{b,z}$</td>
<td><strong>0.02070</strong> $^{b,z}$</td>
</tr>
<tr>
<td>Pareto/NBD</td>
<td>Exponential: $s$</td>
<td>0.04929</td>
<td>0.04536</td>
<td>0.03982</td>
<td>0.03669</td>
</tr>
<tr>
<td></td>
<td>Exponential: $b,z,s$</td>
<td>Erlang-2: $b,z,x,s$</td>
<td>Erlang-2: $b,z,x,s$</td>
<td>Erlang-2: $b,z,x,s$</td>
<td>Erlang-2: $b,z,x,s$</td>
</tr>
<tr>
<td></td>
<td>Exponential: $b,z,s,z$</td>
<td>Erlang-3: $b,z,x,s,z$</td>
<td>Erlang-3: $b,z,x,s,z$</td>
<td>Erlang-3: $b,z,x,s,z$</td>
<td>Erlang-3: $b,z,x,s,z$</td>
</tr>
</tbody>
</table>

$b$, $z$, $s$, $sz$ indicate statistical significance ($p < 0.05$) compared to the base, zero-inflated, seasonal, and seasonal/zero-inflated models respectively while holding the transaction distribution and seasonal parameter equivalent. For the Pareto/NBD, statistical significance to the other models broken out by distribution.

Now there are multiple elements that affect estimates of $P(\text{Alive})$, not just that we are correctly modeling the zero-repeat customers. As described in Chapter 3, with nonstationary transaction data the summary statistic for the right-censored time period, $E[T - t_n]$, is no longer i.i.d.. We need the seasonal correction to smooth out estimates of $P(\text{Alive})$ over time by adjusting the size of the transaction survival function. The choice of transaction distribution by the analyst also affects the size of survival function of the transaction process has reached over $T - t_n$, which goes on to affect estimates of $P(\text{Alive})$. With Figures 5.11 through 5.13, we can see these three effects of the CLV modeling framework. First, the seasonally corrected models produce estimates less biased by the timing of model fitting. Second, the exponential transaction distribution overestimates $P(\text{Alive})$ while the Erlang-3 distribution underestimates $P(\text{Alive})$. Finally, once we have a good choice for the transaction distribution (Erlang-2) and a seasonal correction, the zero-inflation specification gets us closer to the true value over time, though not
always that exact (see Table 5.3). The Pareto/NBD is included in these figures and Table 5.3. Note that the Pareto/NBD has in most cases the most bias and RMSE with $P(Alive)$ of any of these models as it is hampered by both an exponential transaction distribution and no zero-inflated specification to model an excess of customers with no repeat transactions.

With clean, simulated data, the improvement of the zero-inflation specification may not seem like much, but with messier, empirical data, I will show later in this chapter that it does greatly improve estimates of $P(Alive)$. Part of the problem is that inference for zero-inflated model specification is quite complicated (see Appendix A) in utilizing the ECM algorithm within a Gibbs sampler in its use of data augmentation as well as mean transaction and attrition rates rather than individual rates that are impossible to infer for individuals with no repeat transaction data. So while we do see much reduced bias with estimates of the $E[p]$, there is potentially a variance trade-off going on with our estimates of $P(Alive)$. That said, the zero-inflation specification is an important tool towards having a CLV model reflect the customer behaviors that affect estimates of future customer behavior for managers.

Next, I look at the model estimates more aligned with the transaction process. Figure 5.14 shows the estimates for the six models that include a seasonal covariate, with the mean estimate and the 95% confidence interval across 25 simulations displayed by time period and covariate size. There are a couple of things worth noting here. First, with a seasonal covariate we are modeling a change in the hazard function due to a seasonal dummy variable. Different transaction distributions have different hazard functions. The magnitude of the parameter under the same conditions will be different in size, but not in direction, based on having a different hazard function. The correct distribution is the Erlang-2. But what we see in Figure 5.14 is that the Erlang-2 models recovery of the parameter value show a tendency to slightly underestimate the true parameter value. This is tied to the granularity of the discrete time units of weeks in the simulation study.
Figure 5.14: Recovery of $\gamma$, Exponential, Erlang-2, and Erlang-3 Transactions
Figure 5.14 (Continued)
Next, I examine the recovery of the transaction rate parameter by looking at $E[\lambda]$. To make these results comparable across models, I adjust the estimates such that they are all comparable to the exponential distribution, where $\lambda \sim \text{Gamma}(4, 64)$. This is accomplished by dividing the Erlang-2 mean parameter estimates by 2, and the Erlang-3 estimates by 3. Figures 5.15 through 5.17 that respectively show the estimated mean bias (Figures a), with the 95% confidence intervals, and RMSE (Figures b) for $E[\lambda]$ as calculated for each dataset for the
exponential, Erlang-2, and Erlang-3 transaction distributions. Here, by far the best fitting set of models follows the CLV modeling framework with the exponential transaction distribution, even though the correct distribution is the Erlang-2 distribution. We can at least take comfort that the Pareto/NBD is also a poor fit, and that the use of the seasonal correction and a zero-inflation model specification does improve the consistency of the estimates over time. It is unknown if the misestimates can be tied to the granularity of the time period being used or to issues of inference with the right-censored time portion of the individual likelihood function, or an issue with overestimating the variance of the population distribution. First, the bias increases as the shape parameter of the transaction distribution increases from 1 (exponential) to 3 (Erlang-3). Second, the estimates improve as the seasonal parameter is smaller. The large seasonal parameter of $\gamma = 3$ implies a larger jump in the CDF from one time unit to the next during the season in question.
Figure 5.15a: Recovery of $E[\lambda]$, Exponential Transactions (Bias)
Figure 5.15b: Recovery of $E[\lambda]$, Exponential Transactions (RMSE)
Figure 5.16a: Recovery of $E[\lambda]$, Erlang-2 Transactions (Bias)
Figure 5.16b: Recovery of $E[\lambda]$, Erlang-2 Transactions (RMSE)
Figure 5.17a: Recovery of $E[\lambda]$, Erlang-3 Transactions (Bias)
Figure 5.17b: Recovery of $E[\lambda]$, Erlang-3 Transactions (RMSE)
Table 5.4: Recovery of E[λ], RMSE

<table>
<thead>
<tr>
<th>Transaction Distribution</th>
<th>Parameter by Model</th>
<th>3</th>
<th>1.5</th>
<th>1</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>Base</td>
<td>0.01128 $^{s,sz}$</td>
<td>0.00440 $^{s,ssz}$</td>
<td>0.00204 $^{s,sz}$</td>
<td>0.00199 $^{s,sz}$</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>0.01084 $^{s,sz}$</td>
<td>0.00390 $^{s,ssz}$</td>
<td>0.00165 $^{b,z}$</td>
<td>0.00255 $^{b,ssz}$</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>0.00112 $^{b,z}$</td>
<td>0.00151 $^{b,ssz}$</td>
<td>0.00203 $^{r,ssz}$</td>
<td>0.00187 $^{r,sz}$</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>0.00108 $^{b,z}$</td>
<td>0.00120 $^{b,z,sz}$</td>
<td>0.00163 $^{b,z}$</td>
<td>0.00139 $^{b,z,s}$</td>
</tr>
<tr>
<td>Erlang-2</td>
<td>Base</td>
<td>0.01239 $^{s,sz}$</td>
<td>0.00586 $^{s,ssz}$</td>
<td>0.00355 $^{b,z}$</td>
<td>0.00091 $^{s,sz}$</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>0.01224 $^{s,sz}$</td>
<td>0.00570 $^{b,z,sz}$</td>
<td>0.00335 $^{b,z}$</td>
<td>0.00078 $^{s,sz}$</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>0.00381 $^{b,z,sz}$</td>
<td>0.00350 $^{b,z,sz}$</td>
<td>0.00373 $^{b,ssz}$</td>
<td>0.00311 $^{b,z}$</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>0.00358 $^{b,z,s}$</td>
<td>0.00332 $^{b,z,sz}$</td>
<td>0.00352 $^{s,sz}$</td>
<td>0.00292 $^{b,z}$</td>
</tr>
<tr>
<td>Erlang-3</td>
<td>Base</td>
<td>0.01396 $^{s,sz}$</td>
<td>0.00750 $^{s,ssz}$</td>
<td>0.00518 $^{s,sz}$</td>
<td>0.00214 $^{s,sz}$</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>0.01385 $^{s,sz}$</td>
<td>0.00736 $^{s,ssz}$</td>
<td>0.00509 $^{s,sz}$</td>
<td>0.00198 $^{s,sz}$</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>0.00709 $^{b,z}$</td>
<td>0.00587 $^{b,z}$</td>
<td>0.00572 $^{b,z}$</td>
<td>0.00473 $^{b,z}$</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>0.00695 $^{b,z}$</td>
<td>0.00570 $^{b,z}$</td>
<td>0.00558 $^{b,z}$</td>
<td>0.00464 $^{b,z}$</td>
</tr>
<tr>
<td>Pareto/NBD</td>
<td>Exponential: $b,z,s,sz$</td>
<td>0.00442</td>
<td>0.00220</td>
<td>0.00456</td>
<td>0.00752</td>
</tr>
<tr>
<td></td>
<td>Erlang-2: $b,z,s,sz$</td>
<td>Exponential: $b,z,s,sz$</td>
<td>Exponential: $b,z,s,sz$</td>
<td>Exponential: $b,z,s,sz$</td>
<td>Exponential: $b,z,s,sz$</td>
</tr>
<tr>
<td></td>
<td>Erlang-3: $b,z,s,sz$</td>
<td>Exponential: $b,z,s,sz$</td>
<td>Exponential: $b,z,s,sz$</td>
<td>Exponential: $b,z,s,sz$</td>
<td>Exponential: $b,z,s,sz$</td>
</tr>
</tbody>
</table>

$b, z, s, sz$ indicate statistical significance (p < 0.05) compared to the base, zero-inflated, seasonal, and seasonal/zero-inflated models respectively while holding the transaction distribution and seasonal parameter equivalent. For the Pareto/NBD, statistical significance to the other models broken out by distribution.

As I did with the attrition process, I also want to briefly examine the variance recovered by the model. The simulation mean implied by the true model parameters for the Erlang-2 model is 0.125. The expected value from all of the Erlang-2 model runs where the seasonal parameter equals 3, corrected for seasonality and zero-inflation, is 0.132, an overestimate of about 5.3%, which is reasonably close. Variance is a different story. The simulation variance implied by the true model parameters is 0.001953. The expected value from all of the same Erlang-2 model runs corrected for seasonality and zero-inflation is 0.004263, an overestimate of 218%. The additional variance appears to be mainly due to utilizing the Erlang-2 distribution. The uncorrected Erlang-2 model narrowed down to week 14 over 25 simulated datasets where the seasonal parameter equals 3 has an average variance of 0.005773, an overestimate of 296%. But the same data fitted with the uncorrected model with the exponential transaction process has an average variance of 0.001365, an underestimate of 30.1%. The Pareto/NBD with the same week 14 data is similar.
with an average variance of 0.001168, an underestimate of 40.2%. On other words, all of the
models are challenged in their attempts to recover the correct variance that represents population
heterogeneity.

Again, there are only two CLV simulation studies I am aware of that report the variance
recovered from simulated data. Korkmaz, Fok, and Kuik (2014), who report in one of their two
simulated latent classes the variance of the transaction rate is 23% more. The other is Platzer and
Reutterer (2016) who only report recovered parameters for five of their 160 simulation scenarios.
In the five simulations they report, they generally recover the mean transaction rates, but the
recovered variance has increased by 13% or more. That is with their selectively reported
simulation scenarios. My experience has been that these models tend to overestimate the
variance with simulated data, though the additional variance when we move away from the
exponential distribution in my modeling framework is particularly troubling.

The bias in the mean estimates of the transaction rates goes on to affect the estimates of
future transactions, since these are based on CDF-simulated counts. The amount of future
transactions over the next two years decreases over time as individuals leave. With increased
seasonal transaction activity, we will again see more individuals leaving over the same period of
time. This is shown in Figure 18, particularly when the seasonal parameter equals 3, and future
transactions decline more quickly during the holidays. Figures 5.19 through 5.21 respectively
show the estimated mean bias (Figures a), with the 95% confidence intervals, and RMSE (Figures
b) for future transactions as calculated for each dataset for the exponential, Erlang-2, and Erlang-
3 transaction distributions.
Figure 5.18: Mean Future Transactions
Figure 5.19a: Recovery of Future Transactions, Exponential Transactions (Bias)
Figure 5.19b: Recovery of Future Transactions, Exponential Transactions (RMSE)
Figure 5.20a: Recovery of Future Transactions, Erlang-2 Transactions (Bias)
Figure 5.20b: Recovery of Future Transactions, Erlang-2 Transactions (RMSE)
Figure 5.21a: Recovery of Future Transactions, Erlang-3 Transactions (Bias)
Figure 5.21b: Recovery of Future Transactions, Erlang-3 Transactions (RMSE)
Table 5.5: Recovery of Future Transactions, RMSE

<table>
<thead>
<tr>
<th>Transaction Distribution</th>
<th>Seasonal Parameter by Model</th>
<th>3</th>
<th>1.5</th>
<th>1</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td></td>
<td>202.9 \text{ <strong>z</strong>}</td>
<td>131.0</td>
<td>138.0</td>
<td>154.7</td>
</tr>
<tr>
<td>Zero-inflated</td>
<td></td>
<td>217.6</td>
<td>175.3</td>
<td>129.6</td>
<td>131.5</td>
</tr>
<tr>
<td>Seasonal</td>
<td></td>
<td>160.3 \text{ <strong>b</strong>}</td>
<td>138.8</td>
<td>139.0</td>
<td>135.9</td>
</tr>
<tr>
<td>Seasonal, ZI</td>
<td></td>
<td>138.3 \text{ <strong>b</strong>}</td>
<td>163.8</td>
<td>133.9</td>
<td>134.5</td>
</tr>
<tr>
<td>Erlang-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td></td>
<td>263.7</td>
<td>196.9</td>
<td>234.0 \text{ <strong>z</strong>}</td>
<td>254.3 \text{ <strong>z</strong>}</td>
</tr>
<tr>
<td>Zero-inflated</td>
<td></td>
<td>243.4</td>
<td>168.2</td>
<td>179.6 \text{ <strong>b</strong>}</td>
<td>181.4</td>
</tr>
<tr>
<td>Seasonal</td>
<td></td>
<td>220.8</td>
<td>201.7</td>
<td>233.4 \text{ <strong>z</strong>}</td>
<td>234.2 \text{ <strong>z</strong>}</td>
</tr>
<tr>
<td>Seasonal, ZI</td>
<td></td>
<td>202.5</td>
<td>165.8</td>
<td>179.1 \text{ <strong>b</strong>}</td>
<td>162.4 \text{ <strong>b</strong>}</td>
</tr>
<tr>
<td>Erlang-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td></td>
<td>334.9</td>
<td>282.5</td>
<td>329.7</td>
<td>350.6</td>
</tr>
<tr>
<td>Zero-inflated</td>
<td></td>
<td>314.6</td>
<td>257.0</td>
<td>288.5</td>
<td>290.9</td>
</tr>
<tr>
<td>Seasonal</td>
<td></td>
<td>308.6</td>
<td>293.3</td>
<td>329.5</td>
<td>322.5</td>
</tr>
<tr>
<td>Seasonal, ZI</td>
<td></td>
<td>307.5</td>
<td>265.2</td>
<td>283.2</td>
<td>267.4</td>
</tr>
<tr>
<td>Pareto/NBD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exponential: s, sz</td>
<td></td>
<td>236.1</td>
<td>190.4</td>
<td>157.8</td>
<td>154.1</td>
</tr>
<tr>
<td>Erlang-2: s, sz</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exponential: b, s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Erlang-2: b, s, sz</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exponential: b, z, s, sz</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b, z, s, sz indicate statistical significance (p < 0.05) compared to the base, zero-inflated, seasonal, and seasonal/zero-inflated models respectively while holding the transaction distribution and seasonal parameter equivalent. For the Pareto/NBD, statistical significance to the other models broken out by distribution.

Future transactions are created by combining an estimate of a number of transactions over time by an individual that is modified by multiplying it by $P(\text{Alive})$ to adjust for the probability they may no longer be active. That creates the seasonal pattern we see in the estimates from models without a seasonal correction. As anticipated with the issue of time unit granularity, the models with the Erlang-2 transaction distribution do underestimate future transaction amounts. But we do get more consistent estimates through use of the seasonal correction as well as the zero-inflated model specification. And despite the fact that the exponential transaction model performs best in terms of future transactions, with the seasonal correction and the zero-inflated model specification, both the exponential and Erlang-2 models outperform the Pareto/NBD.

Finally, as another way to compare the different models, I include a plot of the Newton-Raftery estimates (Rossi, Allenby, and McCulloch 2005) with Figure 5.22. This statistic is a
weighted average of the model likelihood over the retained iterations of the MCMC chain. This does not include a penalty for additional parameters or model complexity. In reality, only two additional population parameters are added to the models. We add population parameter $\omega$ within the zero-inflated models, and we have $\gamma$ for the models with the seasonal correction. Otherwise, the models contain the same number of freely estimated individual transaction and attrition rate parameters.

The plots are a mean over the 25 simulations, with the 95% confidence intervals. The estimates decrease with time as more past transaction data becomes available on which to fit models. As the simplest model that does not need to be broken into discrete time units, the Pareto/NBD does perform best here. But not that much better, and it is clear from the previous simulation results that the Pareto/NBD does not produce the best estimates of future customer activity. The Newton-Raftery estimates do favor the exponential distribution over the Erlang-2 and Erlang-3 distributions. This fit statistic is unable to identify the correct transaction distribution with simulated data.
Figure 5.22: Newton-Raftery Estimates, Exponential, Erlang-2, and Erlang-3 Transactions
Figure 5.22 (Continued)
The previous simulation study highlighted the main benefits of the CLV modeling framework. We get more consistent, less biased estimates of $P(\text{Alive})$, as well as estimates of future transactions with less timing bias, if still slightly biased. This is a huge accomplishment. Despite the continuing challenges in this area of modeling highlighted by the simulation study,
these benefits are extremely important for managerial purposes. If our CLV models cannot provide consistent estimates to managers because our model estimates are thrown off by seasonal variation and the number of customers with no repeat transactions, these CLV models have no benefit to managers.

To illustrate how the CLV modeling framework can greatly improve estimates for managers over existing CLV models and despite the ongoing modeling challenges, I apply my CLV modeling framework onto the purchase of soup over time. It bears repeating that soup is a product category well known for its seasonal demand in the United States, with soup in greater demand during times of colder weather. Soup is an interesting product category to examine with my CLV modeling framework because its climate-driven transaction pattern is particularly troublesome with existing CLV models. The transaction pattern is also exogenous in nature since managers cannot control climate.

For the empirical study, I use a 2 x 2 x 2 design with two transaction distributions (exponential / Erlang-2), zero-inflation specification (yes / no), and correction for seasonality (yes / no). Partly, this is to keep the model comparisons more limited to be easier to interpret. There is no reason to include the Pareto/NBD here after the simulation study. Also, the Erlang-3 had trouble converging, which leaves nothing for comparison purposes. For each of the 112 separate model fittings (8 models x 14 time periods), I ran one MCMC chain for 20,000 iterations, discarding the first 10,000 iterations as burn-in. This took around 15 hours to run. I note here, and in Appendix A as well, that the transaction rate parameter can sometimes wander off if the chain is left to run for too long. There is a computational trade-off here. More chains would make for a more stable estimates for comparison, but that would require more computational resources. There is also the strong possibility that with uncertainty in the right-censored portion of the individual likelihood function that the overall model is only weakly identified. The empirical study makes it point in favor of the CLV modeling framework with a seasonal
correction, zero-inflated model specification, and a transaction distribution besides the exponential even with just one chain for each model run.

To model the observed seasonality of transactions in the soup panel data, I utilize a standard cosine trend model. I define two population level covariates, $x_1$ and $x_2$, as

$$
x_1 = \cos(2\pi t / \text{period}) \quad \text{and} \\
x_2 = \sin(2\pi t / \text{period}),
$$

where $t$ is the week since the start of the observation period and the period in weeks is equal to $365.25 / 7$ (approximately 52.18 weeks). $\beta_1$ and $\beta_2$ are the respective parameters to be estimated.

My use of the cosine trend model to provide a seasonal correction is specific to the empirical data being modeled. Different empirical data will have different seasonal patterns. In order to provide consistent estimates, it is up to the researcher or analyst to provide a seasonal correction based on the observed data being modeled. The choice of seasonal correction is an \textit{a posteriori} decision and not \textit{a priori} knowledge. The cosine trend does make for an important demonstration of my modeling framework in that the two covariates change every single time period. This is the extreme case for the computational resources necessary to estimate the framework.

I compare model performance with two summary statistics pooled over the 14 model fittings. The root mean squared error (RMSE) compares model estimates versus the actual aggregate amounts. The mean absolute percent error (MAPE) compares model estimates as a percentage deviation from the actual aggregate values. I use the Makridakis (1993) formulation here, where $\text{APE} = \left| \frac{\text{Actual} - \text{Estimate}}{(\text{Actual} + \text{Estimate})/2} \right|$ for each model fitting period. For all the figures for the empirical study, the 95% confidence intervals at the model fitting times are plotted around the posterior means. The 95% confidence intervals are calculated as 1.96 times the standard error that has been corrected for MCMC chain autocorrelation via the CODA package (Plummer et al. 2006), plus or minus the mean of the estimate in question. The confidence intervals are included
to emphasize the mean model estimates for most of the 14 model fitting time periods are significantly different from each other.

I start by examining the recovery of the proportion of customers with no repeat transactions at the time of model fitting. I calculate $\omega + (1-\omega)E[p]$ for each model fitting time period for the four models with a zero-inflation specification, which I plot in Figure 5.23 along with $E[p]$ for the four models that do not have a zero-inflation specification. I plot the actual proportion of customers with no repeat transactions with a wider, gray line as all four models with the zero-inflated specification effectively overlap the actual proportion. As predicted and as with the previous simulation study, the models without a zero-inflation specification greatly underestimate the proportion of customers with no repeat transactions. Thus, our cohort of soup customers has an excess of zero-repeat customers, and the zero-inflation model specification is appropriate.

Figure 5.23: Estimated Proportion of Zero-Repeat Soup Customers at Model Fitting
Next, I look at the estimates of $P(Alive)$ compared to the actual proportion in Figure 5.24, with the models’ respective RMSE and MAPE in Table 5.6. Here we see the importance of the zero-inflation specification compared to the actual amount far more than with clean simulation data. The models without a zero-inflation specification tend to overestimate the mean probability that customers remain active by sizeable amounts of around 0.1 or more. Overestimating the proportion of active customers by an additional 0.1 is a huge overestimate for managers to work with. Further, models without a seasonal correction predict customers are more likely to be active on average when the models are fitted soon after a period of more transactional activity. This decreases over the course of the year as we fit the models when the end of the observation period is further away in time from that peak period. Again, this is caused by the fact that with nonstationary transaction data the length the right-censored time period, $E[T - t_n]$, is not i.i.d.. In terms of $P(Alive)$, the best fitting model of the eight models by both RMSE and MAPE is the Erlang-2 transaction model with the cosine trend seasonal covariates and the zero-inflation specification. The worst fitting model has an RMSE and MAPE 9 times larger than the best fitting model. As a result, I believe the zero-inflated specification is crucial for improving CLV model estimates of $P(Alive)$ for managers.
Figure 5.24: Estimated Soup E[P(Alive)] at Model Fitting

![Exponential Transaction Process](image1)

![Erlang-2 Transaction Process](image2)

Table 5.6: Estimated Soup E[P(Alive)], RMSE and MAPE

<table>
<thead>
<tr>
<th>Transaction Distribution</th>
<th>Base Model</th>
<th>Zero-Inflated</th>
<th>Seasonal Correction</th>
<th>Seasonal, Zero-Inflated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>RMSE</td>
<td>0.2345</td>
<td>0.1172</td>
<td>0.2336</td>
</tr>
<tr>
<td></td>
<td></td>
<td>e: b, z, s</td>
<td>e: b, z</td>
<td>e: b, z, s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2: b, z, s</td>
<td>2: b, z</td>
<td>2: b, z, s</td>
</tr>
<tr>
<td></td>
<td>MAPE</td>
<td>33.24%</td>
<td>18.03%</td>
<td>33.47%</td>
</tr>
<tr>
<td>Erlang-2</td>
<td>RMSE</td>
<td>0.1210</td>
<td>0.0441</td>
<td>0.1065</td>
</tr>
<tr>
<td></td>
<td></td>
<td>e: b, s</td>
<td>e: b, z, s</td>
<td>e: b, z</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2: b, z</td>
<td>2: b, z</td>
<td>2: b, z, s</td>
</tr>
<tr>
<td></td>
<td>MAPE</td>
<td>17.22%</td>
<td>5.92%</td>
<td>16.07%</td>
</tr>
</tbody>
</table>

**RMSE Statistical Significance:** b, z, s, sz indicate statistical significance (p < 0.05) compared to the base, zero-inflated, seasonal, and seasonal/zero-inflated models respectively. The e and 2 note whether the comparison is to the exponential or Erlang-2 models.

All of the models with the exponential transaction process provide significant overestimates of P(Alive) compared to the actual values of the overall cohort of soup customers.

The insight of Platzer (2008) and Platzer and Reutterer (2016) is that we need to specify a transaction process for our latent attrition model that more closely resembles the transaction behavior of the customers we are modeling. By using an Erlang-2 transaction process, we get
estimates that are much closer to the actual cohort values for both future transactions over the next 104 weeks and for $P(Alive)$.

Next, I look at estimates of future transactions. The total expected number of future transactions during the 104 weeks that follow the end of the model fitting period for all eight models, are presented graphically in Figure 5.25, with the RMSE and MAPE presented in Table 5.7. Figure 5.25 also includes the actual total number of future transactions from the data over the next 104 weeks that follow the model estimation time period. The actual data here helps to make salient the timing bias found in estimates from stationary CLV models that are fitted to seasonal transaction patterns.

*Figure 5.25: Forecasted Soup Transactions, 104 Weeks Forward Looking*
Table 5.7: Forecasted Soup Transactions, RMSE and MAPE

<table>
<thead>
<tr>
<th>Transaction Distribution</th>
<th>Base Model</th>
<th>Zero-Inflated</th>
<th>Seasonal Correction</th>
<th>Seasonal, Zero-Inflated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>RMSE</td>
<td>1,629.86</td>
<td>1,431.64</td>
<td>1,591.49</td>
</tr>
<tr>
<td></td>
<td>MAPE</td>
<td>20.43%</td>
<td>17.28%</td>
<td>20.71%</td>
</tr>
<tr>
<td>Erlang-2</td>
<td>RMSE</td>
<td>1,061.53</td>
<td>786.96</td>
<td>928.39</td>
</tr>
<tr>
<td></td>
<td>MAPE</td>
<td>12.62%</td>
<td>8.45%</td>
<td>12.49%</td>
</tr>
</tbody>
</table>

RMSE Statistical Significance: \(b, z, s, sz\) indicate statistical significance (\(p < 0.05\)) compared to the base, zero-inflated, seasonal, and seasonal/zero-inflated models respectively. The \(e\) and \(2\) note whether the comparison is to the exponential or Erlang-2 models.

As predicted, estimates without the seasonal correction anticipate more future transactions when estimated soon after a period of more transactional activity, which then decreases over the course of the year as the models are estimated further away in time from that peak period. At week 9, 2003, the uncorrected exponential transaction model forecasts 10,671 transactions over the next 104 weeks, whereas the actual amount is 8,331, a 28% overestimate. At the same time, the Erlang-2 transaction model with both seasonal and zero-inflated corrections forecasts 8,764 future transactions, an overestimate of only 5.2%. If these estimates had been used by managers to allocate resources, a 5% over-commitment is potentially far less costly after the fact than a 28% over-commitment. We can explore this further with a rough estimate of the two-year future value of the customer base using the average value per transaction of \(~$4.57\). The overestimate from the uncorrected exponential model of 10,671 future transactions corresponds to a valuation of the cohort of \$48,735 whereas the true future number of transactions of 8,331 corresponds to \$38,048. With only 1,775 households, we overestimate the value of the cohort by over \$10,600, or about \$6 per household. With consideration that we are focusing on soup, \$6 is a nontrivial amount. That overestimate could lead managers to considerably misallocate marketing resources.
Now looking at week 33, 2003, the uncorrected exponential transaction model forecast estimates decline 22% to 8,341 future transactions over the following 104 weeks, while the corrected Erlang-2 transaction model forecasts 7,690 future transactions, a decline of 12%. The actual amount is 7,550, a decline of 9.4%. The 22% drop in less than six months may lead managers to believe their existing customers are leaving more quickly than is true. That 22% decline in transactions also corresponds to the value of the cohort dropping more than $10,600. In comparison, the corrected Erlang-2 transaction model forecasts a drop in the value of the cohort of roughly $4,900, or $2.76 per household, compared to an actual drop in value of around $2.01 per household. Again, these poor estimates by the uncorrected model could lead to poor resource allocation decisions.

Overall, the seasonal correction through time-varying covariates provides an estimate less biased by the timing of model fitting. Furthermore, the models without a zero-inflation specification tend to further overestimate the number of future transactions, which is tied to the estimates of $P(\text{Alive})$ that I previously discussed, which is itself tied to $E[T - t_n]$ no longer being i.i.d.. In terms of the estimated future number of transactions, the best fitting model by both RMSE and MAPE is the Erlang-2 transaction model with the cosine trend seasonal covariates and the zero-inflation specification. I note that the largest future transaction RMSE of the eight models is less than the size of the cohort of 1,775 households. Results at the aggregated customer level provide a better view of the bias in latent attrition model estimates without correction, and it is the aggregate estimates that would most influence managerial decision making.

Finally, I examine the parameter estimates for the cosine trend seasonal correction. I plot the estimated parameter values for each of the four models with the seasonal correction for each time period in Figure 5.26. Figure 5.26 includes neither the two baseline models nor the two models with only a zero-inflation specification. We do not have the true parameter values for comparison, but we can compare parameter estimates from the two models with the exponential transaction distribution to a Poisson GLM regression on the 6 years cohort data (Figure 2).
regression provides parameter estimates of $\beta_1 = 0.437$ ($z = 49.41$), and $\beta_2 = -0.158$ ($z = -18.29$). The exponential transaction model parameter estimates are similar.

Figure 5.26: Estimated Cosine Trend Seasonal Correction Parameters

With a proportional hazards formulation for the covariates, we are modeling the change in the cumulative probability of a transaction happening. Different transaction distributions will have parameter values that differ in scale as their respective hazard functions are dissimilar. But the direction, or sign, of the parameter values are expected to remain the same, as the covariate is expected to increase or decrease the hazard encountered.

Discussion

The overall benefit of the general CLV modeling framework is that we can address seasonally biased model estimates through the incorporation of time-varying covariates, address
biased estimates from customers with no repeat transactions through a zero-inflated model specification, while maintaining the flexibility of transaction process specification needed to model transactional data with different distributions. More importantly, with the general latent attrition model framework we can achieve a very close estimate of \( P(Alive) \) so that we have a much more accurate assessment of customer churn in a noncontractual context. Furthermore, I demonstrate this benefit over an extended 104-week model estimation window. The empirical studies in most of the CLV modeling literature predict over weeks or months, not years.

There has been piecemeal usage of time-varying covariates in the CLV literature; I have generalized how time-varying covariates should be incorporated into these models. The zero-inflation model specification is specifically new to latent attrition models, and my results show this is a particularly needed addition to these models to correctly infer the attrition process. And in terms of what estimates we as analysts can provide managers, I have shown that the whole of my framework is greater than the sum of its parts. My framework is an important milestone for the CLV literature.

That said, there is clearly work still needed moving forward. The inability to correctly recover the population variance parameters of both the transaction and attrition processes is problematic. The failure to infer the correct transaction distribution with simulated data remains as a major difficulty. But we should not let these challenges, which we must address in future research, detract from the progress made up to this point.
CHAPTER 6: THREE CHALLENGES EXTENDING THE CLV MODELING FRAMEWORK

While the managerial benefits of the CLV modeling framework are genuine, through the simulation study in Chapter 5 I highlighted ongoing inference challenges with the framework. It is problematic that we cannot identify by statistical fit the correct transaction distribution and that we overestimate the variance of the transaction rate and attrition rate population distributions. In this chapter, I explore three common modeling extensions that have seen either infrequent or minimal use in the CLV modeling literature, that also prove to be problematic with the CLV modeling framework. These three modeling extensions are 1) using the Weibull distribution for modeling the transaction process, 2) using a bivariate lognormal/logitnormal prior for the transaction and attrition processes, and 3) a latent class CLV model of active customers. I have not been able to recover the shape parameter of the Weibull distribution with simulated data. In fact, there is a strong downwards bias towards 0 for the Weibull shape parameter due to the individual likelihood function. With the bivariate prior, we recover a positive, nontrivial correlation between the transaction and attrition rate parameters, even though the simulated processes are independent, which could lead to misinterpretation by many analysts. And with the latent class model, there is a discriminant analysis problem of where by observed transaction data alone we are unable to correctly assign an individual to the right class.

To examine the first two cases, I simulate data with an exponential transaction process and geometric attrition process, with the transaction and attrition rate parameters generated independently. There is no covariate affecting either process. For each set of data generated, there were 1,000 individuals, with the transaction and attrition rate parameters distributed as follows:

\[ \lambda \sim \text{Gamma}(4, 64), \text{ such that } E[\lambda] = 1/16, \text{ or } 0.0625, \text{ with } SD[\lambda] = 0.03125, \text{ and} \]

\[ p \sim \text{Beta}(8, 160), \text{ such that } E[p] = 1/21, \text{ or } \sim 0.0476, \text{ with } SD[p] \sim 0.0164. \]
These are the same distributions as I had used in the simulation study in the previous chapter. There was no simulated excess of customers with no repeat transactions.

The exponential distribution is equivalent to the Weibull distribution with a shape parameter equal to 1. The transaction data was generated as described in more detail in the simulation study in Chapter 5. Figure 6.1 shows an example of the generated transaction data over time.

![Simulated Exponential Transactions](image)

*Figure 6.1: Simulated Transactions Example, Exponential Transaction Process*

I use the exponential distribution for the transaction process, with no covariates, as this represents the simplest set of data on which to fit a CLV model. With both the Weibull and lognormal/logitnormal prior cases, I only examine the data after the equivalent of 4 years of data have been collected. I use the combination of simple models with a large amount of observed simulation data to give the Weibull and bivariate prior cases the best possible opportunity to prove their worth within CLV models.
**The Case of the Weibull Distribution**

The PDF of the Weibull distribution is \( f(t) = \lambda a(\lambda t)^{a-1} \exp(-\lambda t^a) \), with CDF \( F(t) = 1 - \exp(-\lambda t^a) \), and baseline hazard function \( \lambda_0(t) = \lambda a(\lambda t)^{a-1} \). The continuous PDF expressed in terms of the hazard function is

(6.1) \( f(t) = \lambda a(\lambda t)^{a-1} \exp \left( -\int_0^t \lambda a(\lambda u)^{a-1} du \right) \).

Taking the integral in Equation 6.1 gives us the PDF initially described. Now if we incorporate Equation 6.1 into Equation 3.4 to bring it into the context of CLV modeling, we get the individual likelihood function

(6.2) \( L(f(t), p|T, n) = \left[ \prod_{j=1}^n \lambda a(\lambda t_j)^{a-1} \exp \left[ - \sum_{j=1}^n \sum_{k=(t_{j-1}+1)}^{t_j} \lambda a \left( \lambda (t_k - t_{j-1}) \right)^{a-1} \right] \right] \times \)

\( (1-p)^n \left[ p + (1-p) \exp \left[ - \sum_{k=t_{n+1}}^T \lambda a(\lambda t_k)^{a-1} \right] \right], \quad n = 0, 1, 2, ... \)

The log-likelihood is then

(6.3) \( \log L(f(t), p|T, n) = \left[ \left[ \sum_{j=1}^n \log(\lambda) + \log(a) + (a - 1) \log(\lambda t_j) \right] + \right. \)

\( \left[ - \sum_{j=1}^n \sum_{k=(t_{j-1}+1)}^{t_j} \lambda a \left( \lambda (t_k - t_{j-1}) \right)^{a-1} \right] \] + \)

\( n \log(1-p) + \log \left[ p + (1-p) \exp \left[ - \sum_{k=t_{n+1}}^T \lambda a(\lambda t_k)^{a-1} \right] \right], \quad n = 0, 1, 2, ... \)

The last portion of the log-likelihood that represents the right-censored time period, \( \log \left[ p + (1-p) \exp \left[ - \sum_{k=t_{n+1}}^T \lambda a(\lambda t_k)^{a-1} \right] \right], \) is particularly problematic. This maximizes when the summation equals 0, which is easily achieved as \( a \) approaches 0. We would hope that the observed data portion of the log-likelihood would counteract this, but it does not with the uncertainty of having the attrition process jointly modeled with the transaction process. Now \( a \) will not equal zero with the separate \( \log(a) \) in the likelihood function, but it can get very close. Just consider that \( \log(1 \times 10^{-10}) \approx -23 \). Thus, \( a \) can become very small while balancing against the right-censored time period.
I have never been able to successfully recover the Weibull shape parameter in a CLV model, so what I offer is two plots in Figure 6.2 of different MCMC chains which attempt to estimate the shape parameter. The top plot has a starting value of 0.5 for the shape parameter. It wanders close to 1, but never gets there, and then soon plunges close to 0, where it remains. The lower plot has a starting value of 1.5 for the shape parameter. It first wanders above 1, then just under 1, and then similarly plunges close to 0, where it remains. This is typical of what I have encountered.

Figure 6.2: Weibull Shape Parameter Example MCMC Chains

There has been one use of the Weibull distribution for the transaction process in the CLV literature, Knox and van Oest (2014). Rather than have population heterogeneity represented by a
probability distribution, the model parameters only vary by latent class. Individual differences are based on the individual covariate data, such as complaints made. This allows for far less freedom for the other parameters to wander about. However, when a hierarchical structure of distributions is used, as has been near universal in the CLV modeling literature, I have demonstrated that the Weibull distribution becomes much more challenging with the existing formulation of the structure of the individual likelihood function. This issue would extend to the 3-parameter generalized Gamma distribution as well.

The Case of the Bivariate Lognormal/Logitnormal Prior

A bivariate prior is not new to the CLV modeling literature, having been first proposed and empirically tested by Abe (2009) and more recently extended into a mixture of normals prior by Korkmaz, Fok, and Kuik (2014). I have now had some time to work with a bivariate lognormal/logitnormal prior for the rate parameters of the individual transaction and attrition processes with simulated data. I have identified some issues with using this bivariate distribution as a prior distribution within a CLV model. First, and most importantly, the prior recovers a positive, nontrivial correlation between the rate parameters of the two processes as customers with higher transaction and attrition rates die off first, even though the generation processes are independent. Second, and of much less significance, using this prior can depend on a somewhat informative prior distribution, more so as the transaction process becomes more complex. And using the bivariate prior does not provide noticeably improved predictions for future activity.

The individual likelihood with an exponential transaction process, geometric attrition process, no covariates, the bivariate normal prior, and the log and logit Jacobian transformations is as follows:
\[ L(f(t), p| T, n) = [\lambda^n \exp(-\lambda t_n)(1 - p)^n] \times [p + (1 - p) \exp(-\lambda(T - t_n))] \times \]
\[ \left[ \frac{1}{2\pi} |\Sigma|^{-1/2} \exp \left( -\frac{1}{2} (\Theta - \mu) \Sigma^{-1} (\Theta - \mu) - \frac{1}{2} \left( \frac{1}{\lambda - p^2} \right) \right) \right], \Theta = \begin{bmatrix} \log(\lambda) \\ \logit(p) \end{bmatrix}, n = 0, 1, 2, \ldots \]

For the hyperpriors, I use the common bivariate normal, inverse Wishart priors, with the mean vector set to 0, and a set of different priors for the inverse Wishart distribution that I explain below. This formulation has been described in multiple Bayesian statistics textbooks, and I made use of the procedure described by Gelman et al. (2014), pg. 73. Their formulation for the inverse Wishart is:

\[ p(\Sigma|S, \nu) \propto |S|^{\nu/2} |\Sigma|^{-(\nu+k+1)/2} \exp \left( -\frac{1}{2} tr(\Sigma S^{-1}) \right), \]

where \( \nu \) is the degrees of freedom, and \( k \) is the dimensions of \( S \).

Alongside the Jeffreys prior, for the other inverse Wishart priors I use a multiple of the identity matrix. This multiple, \( \nu \), equals the number of dimensions of the prior (2 for the bivariate) plus either 3, 8, 15, so \( \nu = 5, 10, 17 \). I use \( \nu \) as the degrees of freedom for the for the inverse Wishart prior. Besides the Jeffreys prior, where the degrees of freedom equal 1, the hyperprior \( S \) for the inverse Wishart is \( \nu I \). The number of prior measurements is 0. In the MCMC sampler, I draw \( \Sigma|\lambda, p \sim InvW_{\nu_n}(\Lambda_n^{-1}) \), where \( \Lambda_n \) equals the hyperprior \( \nu I \) plus the sum squared matrix about the sample means, and \( \nu_n \) equals 1,000 (the number of individuals) plus \( \nu \). I use the \( rWishart() \) function in R’s base stats package to make this draw. I had no issues with convergence, barring the Jeffreys prior.

I simulated 50 different datasets as described above, each with 1,000 individuals. For the analysis, I first look at the estimated correlation between the transaction and attrition rate parameters, and then I examine the RMSE and MAPE for four model estimates – future transactions over the next two years, \( P(Alive) \), \( E[\lambda] \), and \( E[p] \) – with the mean estimate over the estimates from the 50 different datasets. The MCMC chain for each model/dataset combination was run for 20,000 iterations, with the first 5,000 discarded as burn-in.
First, Table 6.1 shows the mean estimated correlation between the transaction and attrition rate parameters, along with the standard deviation and standard error from the 50 estimates from each model. With the Jeffreys prior, there was no convergence. With the estimated standard deviation, values with the Jeffreys prior model could potentially range above 1 and below -1. What winds up happening is the recovered value for the variance of the attrition process is extremely small, with most of the variance instead estimated as the covariance parameter between the two processes. With the other priors I tested, the mean estimates of the correlation are around 0.1, which are significantly different from 0 (Table 6.1).

Table 6.1: Average Bivariate Normal Model Estimates of Correlation Between Simulated Transaction and Attrition Rate Parameters by Prior

<table>
<thead>
<tr>
<th>Prior</th>
<th>Jeffreys</th>
<th>Dim + 3</th>
<th>Dim + 8</th>
<th>Dim + 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>-0.3965</td>
<td>0.1018</td>
<td>0.1065</td>
<td>0.0920</td>
</tr>
<tr>
<td>SD</td>
<td>0.6906</td>
<td>0.1413</td>
<td>0.1234</td>
<td>0.0820</td>
</tr>
<tr>
<td>SE</td>
<td>0.0977</td>
<td>0.0200</td>
<td>0.0174</td>
<td>0.0116</td>
</tr>
<tr>
<td>t</td>
<td>-4.06**</td>
<td>5.09**</td>
<td>6.12**</td>
<td>7.93**</td>
</tr>
</tbody>
</table>

*Dim = 2; ** p < 0.01*

These are two independent processes. The only reason behind this positive correlation is that individuals with higher transaction rates and higher attrition rates will on average leave more quickly. If this were a small correlation, say, \(|p| < 0.05\), we could potentially dismiss it as meaningless even when it is statistically significant. But the correlation is sizeable enough that most analysts would not dismiss it as meaningless. And that is a problem. It means the model estimates are conveying false information.

Now were the other estimates from the model closer to the true values, there is a possible tradeoff to be made. We could potentially accept a problematic correlation to at least get closer to the truth elsewhere in the model estimates. But this is not the case. Looking at the average estimates of future transactions over the next two years, \(P(\text{Alive})\), \(E[\lambda]\), and \(E[p]\), the
improvement in model estimates is minimal. Tables 6.2 and 6.3 show the RMSE and MAPE respectively of these estimates. Whatever improvement there are to be made, they are not large enough to be worth conveying a false correlation.

Table 6.2: RMSE of Average Exponential and Bivariate Normal Model Estimates

<table>
<thead>
<tr>
<th>Model / Prior</th>
<th>Exponential</th>
<th>Jeffreys</th>
<th>Dim + 3</th>
<th>Dim + 8</th>
<th>Dim + 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Future Transactions</td>
<td>123.5</td>
<td>300.6</td>
<td>126.9</td>
<td>119.2</td>
<td><strong>114.5</strong></td>
</tr>
<tr>
<td>$P(Alive)$</td>
<td>0.0147</td>
<td>0.0145</td>
<td>0.0142</td>
<td><strong>0.0138</strong></td>
<td>0.0151</td>
</tr>
<tr>
<td>$E[\lambda]$</td>
<td><strong>0.0021</strong></td>
<td>0.0029</td>
<td>0.0036</td>
<td>0.0037</td>
<td>0.0037</td>
</tr>
<tr>
<td>$E[p]$</td>
<td>0.0070</td>
<td><strong>0.0035</strong></td>
<td>0.0059</td>
<td>0.0071</td>
<td>0.0076</td>
</tr>
</tbody>
</table>

$Dim = 2$

Table 6.3: MAPE of Average Exponential and Bivariate Normal Model Estimates

<table>
<thead>
<tr>
<th>Model / Prior</th>
<th>Exponential</th>
<th>Jeffreys</th>
<th>Dim + 3</th>
<th>Dim + 8</th>
<th>Dim + 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Future Transactions</td>
<td>4.00%</td>
<td>11.18%</td>
<td>4.05%</td>
<td>3.79%</td>
<td><strong>3.46%</strong></td>
</tr>
<tr>
<td>$P(Alive)$</td>
<td>2.14%</td>
<td>2.24%</td>
<td>2.14%</td>
<td><strong>2.11%</strong></td>
<td>2.25%</td>
</tr>
<tr>
<td>$E[\lambda]$</td>
<td><strong>3.08%</strong></td>
<td>4.28%</td>
<td>5.34%</td>
<td>5.54%</td>
<td>5.55%</td>
</tr>
<tr>
<td>$E[p]$</td>
<td>12.24%</td>
<td><strong>5.11%</strong></td>
<td>9.97%</td>
<td>11.81%</td>
<td>12.85%</td>
</tr>
</tbody>
</table>

$Dim = 2$

I will only touch on the last issue with the bivariate prior, which I consider minor. When covariates are added to the transaction process, a stronger prior on the inverse Wishart hyperprior distribution is necessary to get the model estimates to converge. This is somewhat common with Bayesian modeling by means of an MCMC sampler. My objection here is the attrition process is latent, and any correlation is unobservable. So yes, we can use a stronger prior, but of what sort? Do we use the identity matrix with a larger multiple, thus favoring no correlation when one may in fact exist? Or do we add a prior that includes a covariance that is unobservable? Because the attrition process is latent, there is no observable data on which to base our prior knowledge. A weak prior may be acceptable, but it becomes more questionable with CLV models as a stronger
prior is needed to achieve convergence. Forcing prior information on a model, when the reality is genuinely unknown, is questionable. It is a minor objection, but it should be noted.

**The Case of the Latent Class Model**

The primary truth we hold onto within Marketing as an academic discipline is that consumers are different from one another. Furthermore, accounting for customer heterogeneity is key for a firm to successfully market to its customers. Towards that end, the marketing modeling literature has many different methodologies to account for how customers may vary from one another. I have incorporated population mixing distributions into the CLV modeling framework to account for customer heterogeneity, but that approach does not account for customer behavior where customers cluster together into different behavioral segments. The most common approach is to model these customer segments as latent classes, e.g. Kamakura and Russell (1989), who model different segments of customers making different price-quality tradeoffs, or Bell and Lattin (2000), who model different reference price responses by different customer segments. We cannot directly observe whether a customer is a member of a specific latent class, but statistically we can assign a probability of class membership.

There have been a few uses of latent classes in the CLV modeling literature to account for this form of customer heterogeneity. Venkatesan and Kumar (2004) and Venkatesan, Kumar, and Bohling (2007) allow for different customer segments in their always-a-share CLV model. Schweidel and Knox (2013) and Knox and van Oest (2014) also extend their models with latent classes. However, these four models do not include hierarchical population mixing distributions. Their use of latent classes allows for customers to cluster at fixed parameter points associated with each segment rather than spread over a continuous distribution, which is different from hierarchical NBD-style model formulations that typifies much of the CLV modeling literature. Korkmaz, Fok, and Kuik (2014) provide the one example of a CLV latent class model that
utilizes hierarchical population distributions by extending Abe (2009) through the use of a mixture of bivariate normal distributions to model different customer segments. This allows for multiple modes to represent different customer segments while maintaining the hierarchical population model so that different customers within a segment can be modeled by a continuous distribution. However, the latent class model of Korkmaz and colleagues has similar variance overestimation issues as I discussed with the simulation study in Chapter 5.

One of the major challenges with implementing latent class models is that we are only able to achieve an exact analysis if we are utilizing exponential family distributions and conjugate priors (Robert and Mengersen 2011). This is one of the major benefits of using a mixture of normal distributions model formulation for latent classes within a Bayesian analysis as the various normal distributions are conjugate with each. However, as I discussed in the previous section, it is questionable whether a bivariate normal distribution correctly models the covariance between the transaction and attrition processes of a CLV model. Analysis of models that fall outside the exponential family can proceed using numerical methods, but an exact analysis cannot be counted upon.

In a marketing context, the mixture of normal distributions model is discussed by Rossi, Allenby, and McCulloch (2005). Gelman et al. (2014) have more general examples of other mixture distribution possibilities, and their book informs my implementation of latent classes within a CLV model. What follows is an extension of my CLV modeling framework from Chapter 5 to allow for not just different covariate parameters, but different transaction and attrition population mixing distribution parameters for each latent class. This allows for a wide range of potential customer patterns to be modeled. This section concludes a simulation study to examine the capabilities and limitations of the latent class extension.
The CLV Modeling Framework with Latent Classes

I start with Equation 5.6, the individual likelihood that includes time-varying covariates as well as a zero-inflated correction, reproduced here for convenience:

\begin{equation}
\begin{align*}
P(n = 0) &= \omega + (1 - \omega) \left[ p + (1 - p) \exp\left\{ - \sum_{t=1}^{T} \lambda_0(t_k) \exp(x_k' \beta) \right\} \right], \\
P(n > 0) &= (1 - \omega) \left[ \prod_{t=1}^{T} \lambda_0(t_j) \exp(x_j' \beta) \right] \exp\left\{ - \sum_{t=1}^{T} \lambda_0(t_k) \exp(x_k' \beta) \right\} (1 - p)^n \times \\
&\left[ p + (1 - p) \exp\left\{ - \sum_{t=t_{n+1}}^{T} \lambda_0(t_k) \exp(x_k' \beta) \right\} \right]
\end{align*}
\end{equation}

where the number of observed repeat transactions \( n = 0, 1, 2, \ldots \); \( 0 < \omega < 1 \), \( \lambda_0(t) \) is the baseline hazard function of transaction distribution \( f(\cdot) \), \( p \) is the rate parameter of the geometric the latent attrition process, \( T \) is the total observed time, \( t_j \) and \( t_k \) indicates the time between each repeat transaction \( j \) from 1 to \( n \), or time unit \( k \) from 1 to \( t_n \), respectively (the \( j \) and \( k \) time indexing is described in more depth in Chapter 5), \( t_n \) is the time of the last observed transaction, and \( \beta \) is a vector of population level covariate parameters.

I extend Equation 5.6 to allow for \( M \) discrete latent classes as:

\begin{equation}
\begin{align*}
P(n = 0) &= \omega + (1 - \omega) \sum_{m=1}^{M} \pi_m \left[ p + (1 - p) \exp\left\{ - \sum_{k=1}^{T} \lambda_0(t_k) \exp(x_k' \beta_m) \right\} \right] \times \\
&g(\lambda|\theta_m)h(p|\phi_m), \quad \text{and} \\
P(n > 0) &= (1 - \omega) \sum_{m=1}^{M} \pi_m \left[ \prod_{t=1}^{T} \lambda_0(t_j) \exp(x_j' \beta_m) \right] \times \\
&\left[ \exp\left\{ - \sum_{k=1}^{T} \lambda_0(t_k) \exp(x_k' \beta_m) \right\} (1 - p)^n \right] \times \\
&\left[ p + (1 - p) \exp\left\{ - \sum_{k=t_{n+1}}^{T} \lambda_0(t_k) \exp(x_k' \beta_m) \right\} \right] g(\lambda|\theta_m)h(p|\phi_m), \\
\sum_{m=1}^{M} \pi_m &= 1.
\end{align*}
\end{equation}

First, I introduce \( \pi_m \), the probability an individual belongs to class \( m \). Next, I include of the prior distributions \( g(\lambda|\theta_m) \) and \( h(p|\phi_m) \). I make the assumption here that the transaction process distribution has one individual level parameter that varies, \( \lambda \), as was assumed in the simulation and empirical studies in Chapter 5, and that both the individual transaction and attrition parameters, \( \lambda \) and \( p \), have independent prior distributions \( g(\cdot) \) and \( h(\cdot) \) respectively. These two
prior distributions have their own vector of parameters, $\theta_m$ and $\phi_m$, that vary by class. Third, there is only one zero-class parameter $\omega$ as there is only one zero-inflated latent class. Last, the vector of covariate parameters $\beta_m$ varies by class, and here I maintain the previous assumption that vector $\beta_m$ has a uniform prior distribution of $I$.

The details for the MCMC sampling algorithm to implement Equation 6.6 are detailed in Appendix B, but I do want to highlight a few of the broader aspects. Within each iteration, we first determine a probability for belonging to each class based on the expected value of the parameters that were drawn from the previous iteration of the sampler, the Expectation step. We then use those probabilities to draw which class the individual belongs to in this iteration. That class affects the parameter draws from the various conditional distributions in a Gibbs sampler, the Conditional Maximization steps. I include the two prior distributions $g(\cdot)$ and $h(\cdot)$ in Equation 6.6 as they contain the individual rate parameters $\lambda$ and $p$, which are needed in the Expectation step to determine the probability of the class to which a customer should be classified.

The draw for which individual is in which class is a draw from a multinomial distribution. It is quite common in latent class models to use a Dirichlet prior on these multinomial probabilities. The Dirichlet parameters are determined by a hyperparameter combined with a count of individuals in each class from that iteration. However, with different sets of simulated data where I used a Dirichlet prior, the outcome was to weigh nearly every individual towards a class based on the initial class counts within less than 10 iterations of the MCMC sampler. The issue here is that which class an individual belongs to is rather indeterminate, as I will show in the simulation study below. When the class size of each MCMC iteration is used, the randomly larger class of that iteration overwhelms the other classes. Now strong hyperparameters can compensate for this behavior, but then we would effectively be

---

4 The marginal distributions of the Dirichlet distribution are Beta distributions. Effectively, the Dirichlet is useful for modeling a vector of probabilities that sums to 1.
predetermining the size of each class. As a result, for the simulation study later this chapter, the multinomial class probabilities are calculated as a weight of the class likelihoods (see Appendix B) rather than drawn from a conditional distribution.

There is an additional issue with regards to class identifiability, what is known as the label switching problem, e.g. Rossi, Allenby, and McCulloch (2005), Gelman et al. (2014). For example, we may have a model with two classes and two underlying distributions, $N(5, 2)$ and $N(-1, 1)$. A model with $\pi_1 = 0.3, N_1(5, 2); \pi_2 = 0.7, N_2(-1, 1)$ is the same as a model where $\pi_1 = 0.7, N_1(-1, 1); \pi_2 = 0.3, N_2(5, 2)$. The Dirichlet prior is one possible way to control for this in many modeling scenarios as a prior weight is put on each class to sort observed data towards its proper class proportion. Another method is to rank order one or more parameters, such as the means of the two normal distributions in the example I gave above. Regardless, some detail regarding class structure needs to be imposed for model convergence. With the CLV latent class modeling framework, I recommend a rank order by a covariate parameter where we have some expectation or theory that anticipates different parameter values by segment, e.g. different customer segments will have different seasonal responses or different responses to changes in price. We do not have a marketing theory to draw on for how transaction or attrition rates may vary by different segments of consumers. However, forcing a covariate to rank order by class does take care of the label switching inference problem.

Another issue to consider is the zero-class of customers with no repeat transactions. Membership in the zero-class needs to be determined before we make inferences regarding the different latent classes. I discuss this in more detail in Appendix B.

Finally, we need to determine the number of latent classes by separately fitting models with a different number of classes, and comparing which model is best by some kind of criteria, e.g. each model’s Newton-Raftery score. With the CLV modeling framework, each customer has their own rate and attrition parameters. If we are analyzing a dataset of 500 customers, that is
1,000 individual parameters. We have a handful of population parameters on top of that based on the number of parameters for the hierarchical population transaction and attrition distributions, along with the covariate parameters. Each latent class adds another handful of population parameters. It is important to keep in mind that we may not gain enough additional explanation for the observed transaction data in terms of the model likelihood or an information criterion value when each latent class adds less than one percent more to the parameter count. In other words, there may be nothing gained by using an information criterion like the Bayesian Information Criterion (BIC) which penalizes additional model parameters.

**Simulation Study**

Because our observed data is transactional, and regardless of latent class membership, there is always a probability of a transaction happening, however small or large a probability. That makes it difficult for the model to distinguish class membership for individuals. When the underlying classes represent very distinct behaviors, I will show the latent class model has been able to make somewhat reasonable, if biased, inferences of the covariate parameters for each class. Yet even with distinct behaviors, the latent class model is unable to recover the correct simulated class sizes. Furthermore, the CLV latent class model does not appear to provide much improvement in terms of the aggregate predictions of future transactions and \( P(Alive) \).

To work within this limitation, I pretested a few different simulation study designs to demonstrate the CLV latent class model before finalizing the study below. For the study design herein, I simulated 50 datasets of 1,000 individuals. 600 belong to a class of infrequently transacting customers, except for a holiday time period when transactions from this group are much more common. 400 belong to another class with frequently transacting customers, who only slightly increase their baseline behavior during the same holiday period. This closely matches the customer group dynamics described by Goodhardt and Ehrenberg (1967), and is a
fairly common marketing segmentation scheme. I simulated the individual transaction rates with latent class population distributions that had narrow variances to make these two groups reasonably distinct from each other.

Except for incorporating two groups instead of one, the design of the latent class simulation is similar to the simulation study in Chapter 5. Transactions are simulated with an exponential distribution, rather than the Erlang-2 I utilized in Chapter 5. To represent an excess number of zero-repeat transaction customers, there is a 10% chance that an individual will leave before they even transact. The individual transaction and attrition rates are drawn from the following distributions to represent population heterogeneity:

**Class One, n = 600:** \( \lambda \sim \text{Gamma}(15, 640) \), such that \( E[\lambda] \approx 0.02344 \), with \( SD[\lambda] \approx 0.00605 \), with holiday covariate \( g_1 = 5 \),

**Class Two, n = 400:** \( \lambda \sim \text{Gamma}(50, 640) \), such that \( E[\lambda] \approx 0.07813 \), with \( SD[\lambda] \approx 0.01105 \), with holiday covariate \( g_2 = 1.5 \), and for both classes,

\[ p \sim \text{Beta}(8, 160) \], such that \( E[p] = 1/21 \), or \( \approx 0.0476 \), with \( SD[p] \approx 0.0164 \).

All other details for simulating data, e.g. the 8-week holiday period, are the same as detailed in the simulation study in Chapter 5.

For the latent class CLV model, I assume the following changes from the Chapter 5 simulation:

\[ \lambda \sim \text{Gamma}(r_m, a_m) \],

\[ p \sim \text{Beta}(s_m, b_m) \],

\( r_m, a_m, s_m, \) and \( b_m \) are separately distributed Gamma with noninformative priors, and

\( m \) is one of \( M \) classes, the count predetermined at model fitting.

With each dataset, I fit three separate models with one, two, or three latent classes. All three models have the seasonal covariate and a zero-inflated class included. I only fit the models at one point in time where we have the equivalent of four years of observed data. I ran one MCMC
chain for each model. Each chain was run for 15,000 iterations, with the first 5,000 discarded as burn-in. The zero-class parameter was correctly inferred by all three models over the 50 simulated datasets, and I will not discuss those results further.

Simulation Study Results

First, I look at whether the CLV latent class model can identify the two-class model as a better fit than the one-class or three-class models based on each model’s Newton-Raftery score (Rossi, Allenby, and McCulloch 2005) for that simulated dataset. Because each dataset has an unrelated fit to the other datasets, I compare the difference between the model scores in proportion to the two-class model’s Newton-Raftery score. I display the proportions graphically in Figure 6.3. The histogram on the left shows the score differences between the one-class and two-class models, while the histogram on the right shows the score differences between the three-class and two-class models. The one-class models have scores that are approximately 1.3% smaller on average while the three-class models have scores that are a little under 0.5% on average.
The CLV latent class model is able to distinguish the two-class model as the best fitting model of the three models with these 50 simulated datasets. However, we needed distinct groups of simulated data to reach this outcome. In some of my pretesting with less distinct simulated data, the model favored more groups over fewer groups as the additional class parameters provided better fit. I also note that the improvement in the Newton-Raftery scores is less than the score declines of proportionally 2-3% when comparing the model fits of the exponential and Erlang-2 transaction models in the simulation study in Chapter 5. I would recommend that the better fitting transaction distribution in terms of future transactions and $P(Alive)$ should be identified before fitting more than one latent class.

Next, I look at how well the CLV latent class model is able to recover the size of the two classes, and here the results are more problematic. With the two-class model, Figure 6.4 shows the size of Class One in proportion of the total number of customers not in the zero class. While
the 600/400 absolute count does not hold because we have approximately 10% of customers with no repeat transactions, the recovered proportion should closely match the original proportion of 0.6. However, Figure 6.4 shows this is centered around 0.516, far off from 0.6. The CLV latent class model has a difficult time distinguishing customers between the two classes, even though the underlying population transaction rate distributions of the simulation are distinct.

Figure 6.4: Proportional Size of Class One, Two-Class Model

This becomes apparent when we look at the implied means and standard deviations for these two hierarchical transaction distributions from their parameters. I focus on the means and standard deviations rather than the parameters of the gamma distributions proper as the statistics are a better comparison to what we are observing in the model fitting results. Class One was simulated such that the mean transaction rate was around 0.02344, with a standard deviation of around 0.00605. The implied values for Class One are 0.0268 for the mean, which is reasonably close, but 0.01231 for the standard deviation. The fitted standard deviation is much larger. The
same holds for Class Two, which was simulated such that the mean transaction rate was around 0.07813, with a standard deviation of around 0.01105. The implied values for Class Two are 0.06902 for the mean, which is not as close but not so far off, yet 0.0256 for the standard deviation. The fitted standard deviation is again much larger. As we encountered in the simulation study in Chapter 5, we continue to overestimate the variance of the population, or class, rate distributions. The underlying challenge in applying latent class modeling to hierarchical transactional models that allow for variance in the transaction rates amongst individuals is that the observed transactional behavior is not as distinct as the underlying generating distributions. The imprecision in determining who belongs to what class goes on to affect the inference of other parameters and model output values. And this is exacerbated by the fact that we are not able to correctly recover the variance of the transaction and attrition population distributions.

I next look at the fitted covariate parameter values with Figure 6.5. The fitted values are 5.235 on average for Class One, compared to the simulated value of 5, and 1.558 on average for Class Two compared to the simulated value of 1.5. These are upwardly biased, though not far from their true values. So, assuming we can identify the correct number of latent classes for the model, we can approximately recover the covariate parameter values, should those values be the focus of interest to the analyst or researcher, if some amount of bias is acceptable.
The value of the CLV latent class model for managerial purposes is more debatable. As in the simulation study in Chapter 5, I look forward two years after the model fitting time period for the actual number of future transactions to compare against the estimates from the three models. Customers without transactions in these two years are considered inactive for model comparison purposes. I focus on the proportional size of the bias between the estimated and the actual values.

Figure 6.6 shows the distribution of the proportional size of the bias for estimates of \( P(\text{Alive}) \). The average proportional bias increases from 4.6% for the one-class model to 6.3% with the three-class model. As we add to the number of latent classes, we increase the bias in the model estimates of \( P(\text{Alive}) \). This goes on to affect the model estimates of future transactions. I demonstrated in Chapter 5 that there is a slight negative bias in estimating future transactions because of the size of the discrete time periods used by the model. We can see this in Figure 6.7
where average proportional bias increases from -5.5% for the one-class model up to -2.3% with the three-class model. Now, from the managerial perspective, there is a trade-off that could be made here. While P(Alive) may be overestimated as the number of latent classes in the model is increased, that reduces the bias in the number of estimated future transactions. And regardless, we are not recovering the true simulated values.
Figure 6.6: Estimated $P(\text{Alive})$ in Proportion to Actual, by Model

- One-Class Model Compared With Actual, $P(\text{Alive})$ Proportion
  - Average: 0.0459

- Two-Class Model Compared With Actual, $P(\text{Alive})$ Proportion
  - Average: 0.0584

- Three-Class Model Compared With Actual, $P(\text{Alive})$ Proportion
  - Average: 0.0629
Discussion of the Latent Class Model Simulation Study

Again, assuming we can identify the correct number of latent classes with the CLV latent class model, we are able to somewhat reasonably recover the covariate parameter values of each latent class, with some amount of bias. For both the manager and the researcher, this might help
us to better understand how different customer segments react to different market conditions in terms of their longer term transactional and retention behaviors. Additionally, the CLV latent class model developed here maintains the hierarchical population distributions that for decades have typified the transaction and CLV modeling literatures. This is an important methodological contribution to the CLV modeling literature.

However, this comes with a handful of caveats that does narrow the scope of that contribution. Primarily, the CLV latent class model does not recover the correct size of the different latent classes. There are multiple parts of the model that affect the probability of when an observed transaction takes place, which makes it difficult for the model to distinguish which customer belongs to which class. We are further hampered by the fact that we continue to overestimate the variance of the population distributions of the transaction and attrition processes. The separate hierarchical distributions for each class, even though they have deep support in the literature and were also specifically utilized to generate the simulated data, complicates the recovery of the underlying parameters. We can see this by tracing the assumed chain of events in CLV models that leads to our observed transactions. The underlying distribution for transaction rates, Gamma($r, a$), could in reality be a very narrow distribution. But we do get a range of various individual transaction rates $\lambda$ from that distribution. Those transaction rates then generate a handful of observed repeat transactions each (with Class Two, the mean rate of 0.078 over 52 weeks is slightly more than 4) from which we attempt to infer the value of $\lambda$ for each individual. That is a very limited number of observations that are being utilized to generate the model estimates of $\lambda$. In plain words, it is difficult to distinguish from a handful of observed transactions whether an individual belongs to one class over another, further impaired by the high variance recovered for each class’s hierarchical population transaction distribution from the fitted models. And that is why it is difficult for the model to distinguish between latent classes, even when the model itself directly replicates the hierarchical assumptions of the simulation study.
That limits our ability to improve predictions of $P(Alive)$ and future transactions with the CLV latent class model as compared to a single-class model. So from a managerial perspective, it is debatable if the additional model overhead is worthwhile as aggregate improvements upon the one-class model by a multiclass CLV model are small enough that they are not likely to affect managerial decision making. Additionally, we are unable to recover unbiased covariate parameters with the latent class CLV model. That is particularly problematic if we want to incorporate consumer behaviors into a CLV model when we expect there to be distinct customer segments, e.g. with reference pricing. At best, we can hope for interesting results, but it would be questionable to draw any strong conclusions.
CHAPTER 7: CONCLUSION

This thesis makes two important contributions to the CLV modeling literature. First, I examine how the general structure of the individual likelihood for CLV models leads to biased estimates when fitted to nonstationary, particularly seasonal, data as well as when fitted to data with an excess of customers with no repeat transactions. I emphasize this bias is intrinsic to the basic structure of CLV models with individual likelihood functions similar to the Pareto/NBD, with this category of models more recently termed latent attrition models. Second, I propose a generalized CLV modeling framework to directly address this inherent bias.

This modeling framework consists of three elements. First, it incorporates time-varying covariates through a proportional hazards formulation to address biased estimates from nonstationary data. Second, it includes a zero-inflation specification that to date has not been used in the CLV modeling literature to address the bias from zero-repeat customers. Third, the framework is flexible enough to accommodate most transaction process distributions. My framework provides better estimates of the proportion of customers likely to remain active as well as estimates of future transactions closer to the true customer values. I applied the general CLV model framework in both a simulation study and an empirical study with transactional data of soup purchases that exhibited sinusoidal nonstationary rates over time. Furthermore, I demonstrated that all three elements of the framework – a seasonal correction through time-varying covariates, a zero-inflation specification, and the transaction process distribution – are necessary to generate model estimates that are close to the true values of nonstationary customer transactional behavior in a noncontractual context where we do expect customers to eventually leave.

The better fitting transaction process in the empirical study was the Erlang-2 distribution. I grant there could be further potential improvements to be gained by using a different transaction distribution. The simulation study in Chapter 5 highlights that we do not yet have a method to
identify the best transaction distribution. That highlights some of the remaining statistical inference challenges with the framework presented here and CLV models in general. First, by using a proportional hazards formulation of the transaction PDF, we also need the CDF of the transaction distribution. The CDF of the Gamma distribution does not have a closed form for non-integer shape parameters, which requires significant additional computational resources. This was a key motivation for applying the Erlang-2 distribution in Chapter 5. When I have tested a model that used the 2-parameter Gamma distribution with simulation data, it required approximately two orders of magnitude more (i.e. 100x) computational time. That said, prior literature on the distribution of time between transactions favors probability distributions such as the Box-Cox hazard (Jain and Vilcassim 1991, Helsen and Schmittlein 1993), log-logistic (Chintagunta and Haldar 1998), generalized Gamma (Allenby, Leone, and Jen 1999), and expo-power (Seetharaman and Chintagunta 2003). Excepting the generalized Gamma, all of these distributions have closed forms for their hazard functions. And unlike the 2-parameter Gamma distribution they all potentially have nonmonotonic hazard functions. Exploring which transaction distribution(s) are best for which product category, as applied to latent attrition models, is an opportunity for future research.

There is an additional challenge if the Gamma shape parameter is inferred rather than fixed that applies to all latent attrition models. I have found with simulated nonstationary data that inference of the shape parameter is biased to be smaller than the true value, even more so for the second shape parameter of the generalized Gamma distribution. When the shape parameter of the 2-parameter Gamma distribution is unknown, Wheat and Morrison (1990b) propose a statistic which uses the first two transaction times to infer the parameter value. A statistic of 1 implies the exponential distribution, 2 implies the Erlang-2 distribution, etc. This would allow us to impute a fixed value for the shape parameter. However, their statistic is dependent on the stationary transaction assumption, and is therefore not applicable to nonstationary transactional data. I calculated a statistic of 0.98 with their method using the first two transaction times for each
customer with 3 or more transactions in the soup purchase data, which implies the exponential distribution would be the correct transactional distribution for the dataset. The empirical results convey otherwise, but that is likely due to the fact that the empirical soup data is nonstationary. There are further challenges with the transaction distribution. Chapter 6 highlighted that with the current formulation of the individual likelihood function, we are presently unable to solve for a model with transaction distributions such as the Weibull distribution.

There is an ongoing challenge of correctly estimating the variance of the population distributions for the transaction and attrition processes, which appears to be impacted by choosing a transaction process more complex than the exponential distribution. This additional variance greatly limits the capabilities of a latent class extension of the CLV modeling framework. Future research will need to come up with alternative methods to approach inference of the likelihood function if we are to incorporate more complex transaction distributions and improve the functionality of the latent class extension.

With that said, I did demonstrate in my simulation study in Chapter 5 that with only one active class, we can reasonably recover a seasonal covariate. And, in the empirical study in the same chapter, the exponential model recovered the same cosine trend covariate parameters as a Poisson GLM regression. I did demonstrate that covariate parameter recovery is biased with a latent class model that has separate population distributions by class, which does currently limit extending my CLV modeling framework into consumer behavior applications. However, we now have time-varying covariates for a single active class that greatly improves estimates for managers of a seasonally transacting customer base. This is a very important achievement for CLV modeling.

My implementation of the zero-inflation specification is an asymmetric downwards correction, where for customers in the zero-state, we assume $P(Alive) = 0$. That assumption has allowed us to get much closer to the true proportion of customers active in a customer base than without the zero-inflation specification where $P(Alive)$ is often overestimated. But the zero-
inflation specification does not impact the specification of the transaction distribution, which could leave us with underestimates of future transactions. Even so, the general latent attrition model framework provides better estimates of the customer behaviors that constitute calculations of customer lifetime value with nonstationary data. In particular, the zero-inflation specification gets us much closer to the average value of $P(\text{Alive})$, which addresses Wübben and von Wangenheim’s (2008) criticism of latent attrition models. This is another major accomplishment.

Finally, there are additional managerial implications to consider with my CLV modeling framework. With the stationary transaction assumption, latent attrition models can be calibrated on summary statistics of the time of their customers’ last transactions and counts of their customers’ repeat transaction, i.e. recency and frequency. With a nonstationary transaction model, these are no longer sufficient statistics; firms will need to store more detailed information of their customers’ past transactions in order to make use of our modeling framework. That said, while my framework is more computationally and data intensive than existing latent attrition models, the resource allocations within firms affected by managerial decisions can easily be large enough to justify the resources for more precise estimates of future customer behavior.

To close, I emphasize the core contributions of this thesis. Stationary latent attrition models generate biased estimates of future transactions when modeling nonstationary transaction behavior. We can alleviate time-biased estimates through the use of time-varying covariates in a latent attrition model to account for nonstationary transactional patterns, particularly seasonal patterns that are all too common in many categories of noncontractual goods and services. My method of adding time-varying covariates through a proportional hazards formulation is also generalizable to other covariates besides a seasonal correction, which may open up many future research opportunities. Additionally, latent attrition models produce biased estimates when there is an excess of customers with no repeat transactions, and I address that bias with a zero-inflated model specification. Finally, we can incorporate these changes within a modeling framework that allows for different transaction process distributions.
Gupta et al. (2006, pg. 151) posed the challenge that, “it is not clear if CLV and the accompanying models are relevant in consumer product industry.” For firms that have noncontractual customer bases that are nonstationary in their transactional behavior, especially consumer packaged goods consumers, my general CLV modeling framework provides managers with less biased estimates of their customers’ future transaction and retention behavior. Even in light of the remaining statistical challenges, which are sizeable, my framework is a necessary step towards the accurate valuation of noncontractual customers as assets to be correctly valued and properly managed (Bolton, Lemon, and Verhoef 2004).
APPENDIX A: MCMC GIBBS SAMPLING ALGORITHM

For model inference, I utilize a Gibbs sampler that is primarily Metropolis within Gibbs sampling (Rossi, Allenby, and McCulloch 2005) for most parameter estimates. I do not make use of the data augmentation (Tanner and Wong 1987) strategy for latent attrition models proposed by Abe (2009) where draws from $P(Alive)$ are used to simplify the individual likelihood based on whether a customer left or remained during the right-censored time period for that iteration of the MCMC chain. This is out of caution due to issues related to the right-censored time period described in Chapter 3, as $P(Alive)$ is defined by that right-censored time period. Multiple published studies, e.g. Abe (2009), Platzer and Reutterer (2016), have made use of Abe’s procedure with reasonable parameter recovery from simulation studies. Because of the hazard function formulation of the transaction process, draws for the transaction parameters using standard distributions are unlikely outside of the exponential distribution. The use of Metropolis within Gibbs for the individual level parameters requires slightly more time for computation, but is not as computationally intensive as the addition of time-varying covariates. I do make use of data augmentation for posterior inference of whether a customer is in the zero-state or whether the customer remained active at the end of the observation that iteration of the MCMC sampler using $P(Zero-State)$, described below, and $P(Alive)$ respectively for those probability draws.

Below I outline my MCMC algorithm, and I follow the outline with detailed explanations of each step.

1. Set initial values for all individual and population parameters.

2. For each customer $i$.
   a. If $n$ (number of repeat transactions) = 0, draw indicator for zero-state with probability $P(Zero-State)$, else zero-state indicator = 0.
      i. If zero-state indicator = 1, set $P(Alive)$ and the count of future transactions to 0 for this iteration of the MCMC sampler. This customer is excluded from
all population parameter inference, excepting \( \omega \), for this iteration.

b. If zero-state indicator = 0, draw indicator for active at end of observation period with probability \( P(\text{Alive}) \).

c. Draw individual attrition parameter \( p \).

d. Draw individual transaction parameter(s) per the transaction distribution assumed by the researcher, \( \lambda \) in our empirical study of soup customers.

e. Draw individual count of future transactions.

i. If zero-state indicator = 1 or active indicator = 0, future transaction count = 0.

3. Draw population parameters for the transaction and attrition population heterogeneity distributions.

4. Draw covariate parameters \( \beta \).

5. Draw population zero-state parameter \( w \).

6. Repeat steps 2 – 5 until convergence is achieved.

Before I describe each step in detail, I simplify the full conditional densities for some of the parameter draws. I rearrange Equation 5.4 as follows to move the survival functions together, and note that both summations in the survival functions start at \( k = 1 \).

\[
L(n, t, T | f(t), p) = \left[ \prod_{j=1}^{n} \lambda_{0}(t_{j}) \exp(x_{j}' \beta) \right] (1 - p)^{n} \times
\left[ (p) \exp \left[ - \sum_{k=1}^{T} \lambda_{0}(t_{k}) \exp(x_{k}' \beta) \right] + (1 - p) \exp \left[ - \sum_{k=1}^{T} \lambda_{0}(t_{k}) \exp(x_{k}' \beta) \right] \right],
\]

\( n = 0, 1, 2, ... \) (A1)

I define

\[
A = (p) \exp \left[ - \sum_{k=1}^{T} \lambda_{0}(t_{k}) \exp(x_{k}' \beta) \right] + (1 - p) \exp \left[ - \sum_{k=1}^{T} \lambda_{0}(t_{k}) \exp(x_{k}' \beta) \right],
\]

where \( A \) is used within the full conditional distributions for \( \lambda, p \), and vector \( \beta \).

1. For setting the initial parameter values, I have found using different simulated data that the MCMC algorithm consistently converges from various different starting points, but convergence is slow due to the level of autocorrelation in the draws of the
population parameters for the attrition and transaction processes.

a. For an individual customer’s attrition and transaction parameters, I recommend setting a starting value at a rough average value, e.g. 0.06 for all $p$’s, generated by a previous model run or even maximum likelihood estimates of the BG/NBD or MBG/NBD.

b. Starting values for the population parameters of the attrition and transaction processes should also begin with estimates from previous model inference.

c. I have found no issue with draws for $\omega$ converging quickly from any starting value between 0 and 1, and autocorrelation in the parameter chain is generally non-existent.

d. For starting values of the vector of covariate parameters $\mathbf{\beta}$, I caution that since the covariate parameters are exponentiated, any value too high, dependent on the covariate and time covered, can quickly lead to memory overflow issues. I have found no convergence issues by setting the starting values to 0, and autocorrelation in the covariate parameters chain is generally non-existent when parameterizing a seasonal correction.

2a. The probability that a customer is in the zero-state, given that they have zero repeat transactions, is

$$p(n = 0, \omega) = \frac{\omega}{\omega + (1-\omega)f(0)} = \frac{\omega}{\omega + (1-\omega)[\beta^k + (1-\beta)\exp(-\sum_{k=1}^T \hat{\lambda}_0(t_k)\exp(x_k\mathbf{\beta})]'}}, \quad 0 \leq \omega < 1 .$$  

(A3)

I discussed in Chapter 3 in my justification for the zero-inflated specification that model inference will generally seek to make the transaction CDF $F(T)$, more specifically $\hat{\lambda}_0(t)$ within $F(T)$, as small as allowed by the population heterogeneity distribution of the transaction process. This is particularly true with MCMC inference, which then biases inference of the zero-state. Customers with no repeat transactions provide no information about their time between transactions given that
two events could be occurring. I recommend that the average parameter values for \( \bar{\lambda}_0(t) \) and \( \bar{p} \) from the previous MCMC iteration be used to estimate \( P(\text{Zero-State}) \) where the individual data is the length of the right-censored time period. This is similar to the ECM (Expectation – Conditional Maximization) algorithm described by Gelman et al. (2014), though applied within a Gibbs sampler. The substitution of expected values treats customers with no repeat transactions as customers who could have had observable repeat transactions had they been average customers.

2b. \( P(\text{Alive}) \) is described in Equation 3.7.

2c. The full conditional likelihood and log-likelihood for \( p \), where \( A \) is as described above, are

\[
L(p|n, s, b, A) \propto p^{s-1}(1 - p)^{n+b-1}A
\]

\[
\log(L(p|n, s, b, A)) = (s - 1)\log(p) + (n + b - 1)\log(1 - p) + \log(A)
\]

2d. The full conditional likelihood and log-likelihood for the individual rate parameter \( \lambda \) following the assumption the transaction process is a Gamma distribution, are

\[
L(\lambda_0(t)|n, r, a, A, t) \propto \prod_{j=1}^{n} \lambda_0(t_j)A\lambda^{r-1} \exp(-\lambda a)
\]

\[
\log(L(p|n, r, a, A, t)) = \left( \sum_{j=1}^{n} \log \left( \lambda_0(t_j) \right) \right) + \log(A) + (r - 1)\log(\lambda) - \lambda a
\]

2e. A closed form for a count distribution will be rare for the general CLV model framework because the framework applies to any number of distributions for the transaction process. The count of future transactions must be simulated over each individual for each iteration of the MCMC chain from time \( T + 1 \) until the desired end of the forecast period \( t^* \). Simulating counts with a CDF is straightforward. The CDF of the transaction process, \( 1 - \exp \left[ -\sum_{k=t_n+1}^{T+1} \lambda_0(t_k)\exp(x_k'\beta) \right] \), along with estimates of \( p \), are utilized for each MCMC iteration to simulate counts with the following algorithm. I tested this future transaction algorithm with simulated data where the transactions were distributed exponentially. I compared the algorithm
results against draws from the minimum of a nonhomogeneous Poisson process and a geometric attrition process with nearly equivalent results. To begin, I draw \( q \sim \text{Uniform}(0, 1) \), and then iterate through time until either \( t^* \) is reached or the CDF becomes greater than \( q \). This is repeated until \( t^* \) is reached. To start

1. If zero-state indicator = 1 or active indicator = 0, then future transaction count = 0, and exit.
2. Draw whether customer remained active after last observed transaction with probability \( p \). If customer is active, continue. Otherwise exit with a 0 count this iteration.
3. The first transaction needs to be simulated with a truncated uniform distribution
   \[
   \frac{1-q}{1-F(T-t_n)},
   \]
   where \( F(\cdot) \) is the CDF of the transaction process, as most customers have a right-censored time period that needs to be accounted for. Draw \( a \sim \text{Uniform}(0, 1) \), and then calculate \( q \) as
   \[
   1 - a \cdot \left(1 - F(T - t_n)\right).
   \]
4. Iterate forward in time unit steps until we either reach \( t^* \), or the CDF becomes greater than \( q \).
5. If there is a transaction before or at \( t^* \), draw whether customer remained active after last observed transaction with probability \( p \). If customer is active, continue. Otherwise exit with total count this iteration.
6. Draw \( q \) from an untruncated uniform distribution.
7. Repeat steps 4 – 6 until \( t^* \) is reached.

I provide one caution simulating future transactions by CDF related to time unit granularity. In the point process literature, the level of granularity used for their proofs is a time unit in which only one event can occur, e.g. Ross (2010). We can approximate this by choosing a granularity in discrete time in which the change in the transaction CDF from one time unit to the next is small for a typical customer, e.g.
under 5% during a peak seasonal time period, and less for other time periods. If the unit of time is not granular enough, and the simulated transactions may need to be more granular in time than the empirical data being modeled, the algorithm I outline above will underestimate the number of future transactions.

3. Draws for the population heterogeneity distributions are specific to the prior distributions specified. I note with a Gamma prior for the transaction process, and a Beta prior for the attrition process, both parameters of the Beta distribution and the shape parameter of the Gamma distribution have non-conjugate marginal distributions. All zero-state customers must be excluded from these draws.

4. The full conditional likelihood and log-likelihood for the vector of covariate parameters $\beta$, where $N_z$ is the total number of customers not in the zero-state this iteration, $J$ is a vector of the number of repeat transactions for non-zero-state customers, and $A$ is a vector of $A$ for each individual, with $i$ the individual index, are

$$L(\beta | x, N_z, J, A) \propto \prod_{i=1}^{N_z} \left[ \left( \prod_{j=1}^{I_i} \exp(x_j'\beta) \right) A_i \right]$$

$$\log(L(\beta | x, N_z, J, A)) = \sum_{i=1}^{N_z} \left[ \left( \sum_{j=1}^{I_i} x_j'\beta \right) + \log(A_i) \right].$$

5. The draw for $\omega$ is straightforward. $\omega \sim Beta(z + \delta_\omega, N_z + \eta_\omega)$ where $z$ is the sum of zero-state customers this iteration, $N_z$ is the total number of customers not in the zero-state this iteration, and the additions are noninformative priors (Rodrigues 2003).

6. With the Metropolis-Hastings draws, autocorrelation of the parameter chains can be somewhat reduced through tuning the range of the parameter draws, which helps with convergence. I note that an MCMC chain can sometimes wander away from convergence with a randomly very large draw for $\lambda$, and I recommend the use of multiple, shorter chains when fitting one model to alleviate this, resources permitting. The 112 model fittings in the empirical study in chapter 5 took around 15 hours on a year-old MacBook Pro using the described MCMC sampler, where most of that
computational time was due to the inclusion of the cosine trend seasonal correction. I estimate use of the 2-parameter Gamma distribution would require at least two orders of magnitude more computational time to complete the equivalent empirical study as the survival function must be repeatedly recalculated for each time unit in the summation of the hazard function for each individual.
APPENDIX B: MODIFICATIONS TO MCMC GIBBS SAMPLING ALGORITHM FOR LATENT CLASSES

For inference of the latent classes in the CLV latent class model, I note that the zero-class takes precedence over the other latent classes. The vector of population covariate parameters is contained within Equation A3. However, with multiple latent classes, we now have multiple vectors. An expected value for each covariate parameter should be used for the zero-class probability draw.

Next, we need to probabilistically assign a customer to a latent class before proceeding with further parameter inference in each iteration of the MCMC sampler. This step occurs between steps 2a and 2b described in Appendix A. First, we need to calculate a model weight for each class based on the parameters of that class from the previous iteration of the MCMC sampler. Let \( w_m \) equal the individual model likelihood of class \( m \), such that from Equation 6.6:

\[
B1 \quad w_m = \left\{ \prod_{j=1}^{n} \lambda_0(t_j) \exp(x_j' \beta_m) \right\} \exp \left\{ - \sum_{k=1}^{r} \lambda_0(t_k) \exp(x_k' \beta_m) \right\} (1 - p)^n \times \\
\quad \left[ p + (1 - p) \exp \left\{ - \sum_{k=r+1}^{r} \lambda_0(t_k) \exp(x_k' \beta_m) \right\} \right] g(\lambda | \theta_m) h(p | \phi_m). 
\]

We then let \( W_m \) equal the sum of those model class likelihoods:

\[
B2 \quad W_m = \sum_{m=1}^{M} w_m .
\]

From here, I define indicator variable \( z_{im} \) for individual \( i \) belonging to class \( m \). The probability the indicator \( z_{im} \) is 1 is then:

\[
B3 \quad P(z_{im} = 1) = \frac{w_m}{W_m}, \quad \sum_{m=1}^{M} P(z_{im}) = 1 .
\]

This creates a vector of multinomial probabilities from which we can draw class membership for individual \( i \) in each iteration of the MCMC sampler.

Effectively the formulation used by Equations B1 through B3 treats \( \pi_m \) from Equation 6.6, the prior probability of class membership, as \( \frac{w_m}{W_m} \). I discussed this in Chapter 6, but it bears repeating. It is quite common in latent class models to use a Dirichlet prior on the multinomial
probabilities of belonging to each class. In each iteration of the MCMC sampler, the Dirichlet parameters are determined by a hyperparameter combined with a count of individuals in each class from that iteration. However, with different sets of simulated data where I used a Dirichlet prior, the outcome was to weigh nearly every individual towards a class based on the initial class counts within less than 10 iterations of the MCMC sampler. The issue with the CLV latent class model is that which class an individual belongs to is difficult to determine for many cases. When the class size of each MCMC iteration is used, the randomly larger class of that iteration overwhelms the other classes. Now strong hyperparameters for the Dirichlet distribution can compensate for this, but then we would effectively be predetermining the size of each class. The formulation I am using allows for the empirical data to have more weight in determining class membership.

Having determined class membership for individual $i$ in this iteration of the sampler, we can proceed with step 2b onwards using the population parameters of class $m$ to which $i$ belongs. With steps 3 and 4, where we are drawing the values of the population parameters, these steps are repeated for each class, where the individual data utilized is the subset of individuals that belong to that class in that iteration.

A note of caution while implementing this modification to allow for latent classes. With this change in the MCMC sampler, we are modifying individual variables into vectors, and vectors into matrices. Extreme care should be taken while programming and debugging these modifications.
APPENDIX C: MCMC CONVERGENCE ASSESSMENT OF CHAPTER 5

SIMULATION STUDY

I designed the simulation study in Chapter 5 to examine how well the CLV latent attrition modeling framework is able to estimate $P(\text{Alive})$ and future transactions, as well as some of the simulation parameter values. A full discussion of the study’s results is in Chapter 5. However, in order to reduce the computational complexity of the simulation study I use one, fairly short, MCMC chain for each model fitting. In this appendix, I examine how well these MCMC chains are able to converge as a stationary chain.

Each chain is run for 5,000 iterations, with the first 2,000 discarded as burn-in. The true population values are utilized as the starting values of each MCMC chain. There are 13 models and 4 separate seasonal conditions in all to examine within the simulation study. We have the Pareto/NBD, and then for the models with the geometric attrition process we have three transaction distributions (exponential, Erlang-2, and Erlang-3), zero-inflation specification (yes / no), and correction for seasonality (yes / no). For each set of parameters used for the simulation, there are 4,550 model fittings in all (13 models * 14 time periods * 25 datasets). While the average model run is approximately 18 seconds, that is 81,900 seconds for 25 datasets, or ~ 23 hours. This is repeated four times for each of the four seasonal parameter values simulated.

To check chain convergence, I use the Heidelberger and Welch convergence diagnostic from the CODA package (Plummer et al. 2006). This is a two-part test where the chain is first tested for stationarity, followed by a half-width test to determine if the means of the first and second halves of the MCMC chain are significantly different from each other (which should not be the case). In the simulation study, nearly every chain that passed the stationarity test also passed the half-width test. What I report below in this appendix is the proportion of MCMC chains for each model that pass the stationary distribution test related to the estimates reported in
the simulation study. There are six tables in all where each cell in each table summarizes the 14
time periods over 25 datasets, i.e. 350 MCMC chains for that model.

**Table C.1: Convergence Tests Passed for MCMC Draws of \( \omega \)**

<table>
<thead>
<tr>
<th>Transaction Distribution</th>
<th>Seasonal Parameter by Model</th>
<th>3</th>
<th>1.5</th>
<th>1</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>Zero-inflated</td>
<td>64.6%</td>
<td>57.4%</td>
<td>56.9%</td>
<td>54.6%</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>66.0%</td>
<td>57.1%</td>
<td>60.9%</td>
<td>58.0%</td>
</tr>
<tr>
<td>Erlang-2</td>
<td>Zero-inflated</td>
<td>68.6%</td>
<td>65.1%</td>
<td>62.0%</td>
<td>62.9%</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>71.1%</td>
<td>69.4%</td>
<td>63.1%</td>
<td>64.6%</td>
</tr>
<tr>
<td>Erlang-3</td>
<td>Zero-inflated</td>
<td>72.0%</td>
<td>68.3%</td>
<td>68.3%</td>
<td>62.0%</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>70.3%</td>
<td>67.7%</td>
<td>71.7%</td>
<td>66.9%</td>
</tr>
</tbody>
</table>

The model parameter that represents the zero-state did have some trouble achieving
convergence with a short number of chain iterations. However, as I discuss in Chapter 5, the
estimates of the proportion of zero-repeat customers from the zero-inflated models is nearly spot
on. Even if the parameter did not achieve full convergence, the zero-inflated models do go on to
show improved model estimates.
Table C.2: Convergence Tests Passed for MCMC Draws of $E[p]$  

<table>
<thead>
<tr>
<th>Transaction Distribution</th>
<th>Seasonal Parameter by Model</th>
<th>3</th>
<th>1.5</th>
<th>1</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>Base</td>
<td>65.4%</td>
<td>67.4%</td>
<td>70.9%</td>
<td>67.7%</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>94.9%</td>
<td>92.3%</td>
<td>91.4%</td>
<td>92.0%</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>71.7%</td>
<td>66.0%</td>
<td>69.1%</td>
<td>73.4%</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>96.3%</td>
<td>92.9%</td>
<td>92.9%</td>
<td>94.9%</td>
</tr>
<tr>
<td>Erlang-2</td>
<td>Base</td>
<td>69.7%</td>
<td>64.6%</td>
<td>68.6%</td>
<td>71.1%</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>95.1%</td>
<td>94.6%</td>
<td>93.4%</td>
<td>94.6%</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>66.3%</td>
<td>69.7%</td>
<td>66.6%</td>
<td>65.7%</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>96.3%</td>
<td>96.9%</td>
<td>94.3%</td>
<td>96.9%</td>
</tr>
<tr>
<td>Erlang-3</td>
<td>Base</td>
<td>69.4%</td>
<td>68.3%</td>
<td>68.3%</td>
<td>67.7%</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>97.1%</td>
<td>95.4%</td>
<td>95.4%</td>
<td>95.7%</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>73.7%</td>
<td>72.6%</td>
<td>70.3%</td>
<td>70.0%</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>95.7%</td>
<td>96.3%</td>
<td>94.9%</td>
<td>96.9%</td>
</tr>
</tbody>
</table>

The zero-inflated models that have improved model estimates in the study are better able to achieve chain convergence for estimates of $E[p]$. Because the chains are started with the true population parameters, and as discussed in Chapter 3 $E[p]$ is likely to be upwardly biased without the zero-inflation correction, it is not surprising that the models without the zero-inflation correction did not achieve convergence as often since they would be estimating different parameter values.
Table C.3: Convergence Tests Passed for MCMC Draws of $E[P(Alive)]$

<table>
<thead>
<tr>
<th>Transaction Distribution</th>
<th>Seasonal Parameter by Model</th>
<th>3</th>
<th>1.5</th>
<th>1</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>Base</td>
<td>58.9%</td>
<td>57.4%</td>
<td>57.1%</td>
<td>53.1%</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>84.3%</td>
<td>83.4%</td>
<td>85.7%</td>
<td>84.6%</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>56.0%</td>
<td>60.0%</td>
<td>49.1%</td>
<td>50.6%</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>88.3%</td>
<td>84.3%</td>
<td>81.7%</td>
<td>86.9%</td>
</tr>
<tr>
<td>Erlang-2</td>
<td>Base</td>
<td>65.1%</td>
<td>64.3%</td>
<td>64.3%</td>
<td>58.6%</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>92.3%</td>
<td>90.9%</td>
<td>88.6%</td>
<td>90.6%</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>58.6%</td>
<td>59.7%</td>
<td>58.9%</td>
<td>58.9%</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>89.1%</td>
<td>93.4%</td>
<td>88.9%</td>
<td>90.0%</td>
</tr>
<tr>
<td>Erlang-3</td>
<td>Base</td>
<td>66.6%</td>
<td>67.1%</td>
<td>64.9%</td>
<td>65.7%</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>93.1%</td>
<td>91.7%</td>
<td>90.6%</td>
<td>89.4%</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>66.6%</td>
<td>65.7%</td>
<td>61.4%</td>
<td>61.1%</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>90.9%</td>
<td>93.4%</td>
<td>90.3%</td>
<td>91.1%</td>
</tr>
<tr>
<td>Pareto/NBD</td>
<td></td>
<td>84.3%</td>
<td>83.4%</td>
<td>85.7%</td>
<td>84.6%</td>
</tr>
</tbody>
</table>

Similar to estimates of $E[p]$, the zero-inflated models that have improved model estimates in the study are better able to achieve chain convergence for estimates of $E[P(Alive)]$. The Pareto/NBD has good convergence here as its related population parameter estimates had good convergence as well. The Pareto/NBD’s estimates of $P(Alive)$ are still biased as discussed in Chapter 5, and good chain convergence will not eliminate biased model estimates.

Table C.4: Convergence Tests Passed for MCMC Draws of $\gamma$

<table>
<thead>
<tr>
<th>Transaction Distribution</th>
<th>Seasonal Parameter by Model</th>
<th>3</th>
<th>1.5</th>
<th>1</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>Seasonal</td>
<td>57.4%</td>
<td>55.1%</td>
<td>61.7%</td>
<td>80.3%</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>55.1%</td>
<td>58.3%</td>
<td>61.4%</td>
<td>74.3%</td>
</tr>
<tr>
<td>Erlang-2</td>
<td>Seasonal</td>
<td>70.6%</td>
<td>65.7%</td>
<td>66.6%</td>
<td>88.9%</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>67.7%</td>
<td>72.0%</td>
<td>70.3%</td>
<td>88.0%</td>
</tr>
<tr>
<td>Erlang-3</td>
<td>Seasonal</td>
<td>76.3%</td>
<td>72.6%</td>
<td>78.6%</td>
<td>87.7%</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>78.9%</td>
<td>77.4%</td>
<td>83.1%</td>
<td>87.1%</td>
</tr>
</tbody>
</table>
Similar to the zero-state parameter, the seasonal parameter did have some trouble achieving convergence with a short number of chain iterations. But as I discuss in Chapter 5, the estimates of the seasonal parameter are close to the true simulation value.

Table C.5: Convergence Tests Passed for MCMC Draws of $E[\lambda]$

<table>
<thead>
<tr>
<th>Transaction Distribution</th>
<th>Seasonal Parameter by Model</th>
<th>3</th>
<th>1.5</th>
<th>1</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>Base</td>
<td>92.3%</td>
<td>90.0%</td>
<td>89.1%</td>
<td>89.1%</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>92.9%</td>
<td>91.7%</td>
<td>94.6%</td>
<td>93.7%</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>90.3%</td>
<td>89.1%</td>
<td>92.0%</td>
<td>87.4%</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>92.0%</td>
<td>91.7%</td>
<td>92.6%</td>
<td>93.4%</td>
</tr>
<tr>
<td>Erlang-2</td>
<td>Base</td>
<td>92.9%</td>
<td>93.4%</td>
<td>93.1%</td>
<td>86.3%</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>95.4%</td>
<td>96.0%</td>
<td>94.3%</td>
<td>92.9%</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>91.4%</td>
<td>90.9%</td>
<td>92.6%</td>
<td>90.9%</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>94.6%</td>
<td>95.1%</td>
<td>94.9%</td>
<td>95.4%</td>
</tr>
<tr>
<td>Erlang-3</td>
<td>Base</td>
<td>88.6%</td>
<td>88.0%</td>
<td>90.0%</td>
<td>89.7%</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>94.3%</td>
<td>94.9%</td>
<td>94.6%</td>
<td>94.3%</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>88.0%</td>
<td>91.4%</td>
<td>90.6%</td>
<td>87.1%</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>95.7%</td>
<td>95.4%</td>
<td>96.0%</td>
<td>95.4%</td>
</tr>
<tr>
<td>Pareto/NBD</td>
<td></td>
<td>92.9%</td>
<td>91.7%</td>
<td>94.6%</td>
<td>93.7%</td>
</tr>
</tbody>
</table>

All of the models show reasonably good convergence for their estimates of $E[\lambda]$. 
Table C.6: Convergence Tests Passed for MCMC Draws of E[Future Transactions]

<table>
<thead>
<tr>
<th>Transaction Distribution</th>
<th>Seasonal Parameter by Model</th>
<th>3</th>
<th>1.5</th>
<th>1</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>Base</td>
<td>36.0%</td>
<td>37.4%</td>
<td>44.3%</td>
<td>38.9%</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>85.4%</td>
<td>83.7%</td>
<td>83.1%</td>
<td>80.6%</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>37.4%</td>
<td>40.3%</td>
<td>34.6%</td>
<td>40.9%</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>89.4%</td>
<td>80.6%</td>
<td>82.3%</td>
<td>85.1%</td>
</tr>
<tr>
<td>Erlang-2</td>
<td>Base</td>
<td>44.0%</td>
<td>38.0%</td>
<td>38.6%</td>
<td>40.6%</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>85.4%</td>
<td>84.0%</td>
<td>86.9%</td>
<td>82.9%</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>38.0%</td>
<td>44.3%</td>
<td>37.1%</td>
<td>37.4%</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>84.9%</td>
<td>85.4%</td>
<td>83.1%</td>
<td>86.0%</td>
</tr>
<tr>
<td>Erlang-3</td>
<td>Base</td>
<td>38.0%</td>
<td>39.4%</td>
<td>43.7%</td>
<td>40.3%</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated</td>
<td>88.0%</td>
<td>85.4%</td>
<td>85.7%</td>
<td>85.4%</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>43.7%</td>
<td>39.7%</td>
<td>39.1%</td>
<td>39.4%</td>
</tr>
<tr>
<td></td>
<td>Seasonal, ZI</td>
<td>82.0%</td>
<td>86.0%</td>
<td>85.7%</td>
<td>85.1%</td>
</tr>
<tr>
<td>Pareto/NBD</td>
<td></td>
<td>87.1%</td>
<td>82.6%</td>
<td>84.0%</td>
<td>82.0%</td>
</tr>
</tbody>
</table>

Estimates of future transactions are created as a multiple of $P(Alive)$. The convergence issues here for the models without the zero-inflation specification are related to the convergence issues of $P(Alive)$ as seen in Table C.3 above.
REFERENCES


196


Fader, Peter S., Bruce G. S. Hardie, and Subrata Sen (2014), "Stochastic Models of Buyer Behavior," In Russell S. Winer & Scott A. Neslin (Eds.), The History of Marketing Science (pp. 165-205), now publishers Inc.


