TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Wirtschaftswissenschaften

Hierarchical Bayesian Models for Predicting Purchase Behavior in Noncontractual Settings Based on Individual and Cross-Sectional Purchase Patterns

Dissertation

Robin Wünderlich

TECHNISCHE UNIVERSITÄT MÜNCHEN
Fakultät für Wirtschaftswissenschaften
Lehrstuhl für Dienstleistungs- und Technologiemarketing

Hierarchical Bayesian Models for Predicting Purchase Behavior in Noncontractual Settings Based on Individual and Cross-Sectional Purchase Patterns

Robin Wünderlich

Vollständiger Abdruck der von der Fakultät für Wirtschaftswissenschaften der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Wirtschaftswissenschaften (Dr. rer. pol.)

genehmigten Dissertation.

Vorsitzender:
Univ.-Prof. Dr. Joachim Henkel

Prüfer der Dissertation:
1. Univ.-Prof. Dr. Florian von Wangenheim, ETH Zürich / Schweiz
2. Univ.-Prof. Dr. Gunther Friedl

Die Dissertation wurde am 04.08.2015 bei der Technischen Universität München eingereicht und durch die Fakultät für Wirtschaftswissenschaften am 15.10.2015 angenommen.

Short Table of Contents

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AR .... Autoregressive
ARIMA .... Autoregressive Integrated Moving Average
ARMA .... Autoregressive Moving Average
ASPE .... Average Squared Predictor Error
AUC .... Area Under the Curve
BB .... Beta Bernoulli
BG .... Beta Geometric
BGR .... Brooks Gelman Rubin
BIC .... Bayesian Information Criterion
BUGS .... Bayesian Inference Using Gibbs Sampling
BVAR .... Bayesian Vector Autoregressive
CE .... Customer Equity
CLV .... Customer Lifetime Value
COM-Poisson .... Conway-Maxwell-Poisson
CRM .... Customer Relationship Management
CRV .... Customer Referral Value
CSI .... Consumer Style Inventory
CUSAMS .... Customer Asset Management of Services
DIY .... Do it Yourself
DMPT .... Dynamic Model of Purchase Timing
e.g. .... exempli gratia
EC .... Error Correction
et al. .... et alii / et aliter
etc. .... et cetera

exp.	expected
GCP	Gamma Conway-Maxwell-Poisson
GG	Generalized Gamma
HB	Hierarchical Bayes
HMC	Hamilton Monte Carlo
HMDO	Hierarchical Bayesian Model with Drop-Out
HSM	Hierarchical Bayesian Seasonal Effects Model
HSMDO	Hierarchical Bayesian Seasonal Effects Model with Drop-Out
HW	Holt and Winters
i.e.	id est
IGG	Inverse Generalized Gamma
IPT	Interpurchase Time
IS	Information System
IT	Information Technology
IW	Inverted Wishart
JAGS	Just Another Gibbs Sampler
loc	locus of causality
MA	Moving Average
MAD	Mean Absolute Deviation
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MH	Metropolis Hastings
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimation
MSE	Mean Square Error
MVN	Multivariate Normal
NA	Not Available
NBD	Negative Binomial Distribution
ONR	Office of Naval Research
p.	Page

p.d.f.	Probability Density Function
p.m.f.	Probability Mass Function
PDO	Periodic Death Opportunity
POS	Point of Sale
pp.	Pages
R	Statistical Software Environment
RASPE	Rooted Average Squared Predictor Error
RFID	Radio Frequency Identification
RFM	Recency, Frequency, Monetary Value
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
SARIMA	Seasonal Autoregressive Integrated Moving Average
sd	Standard Deviation
SG	Shifted Geometric
URL	Uniform Resource Locator
VAR	Vector Autoregressive
VARIMA	Vector Autoregressive Integrated Moving Average
VARMA	Vector Autoregressive Moving Average
WinBUGS	Windows Software for Bayesian Inference Using Gibbs Sampling

Chapter 1

Introduction

1.1 Challenges in Predicting Individual Customer Behavior

"... as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know."

—Donald H. Rumsfeld[1]

Marketing executives managing the customer base are in a dilemma. Advances in information technology have led to the accumulation of large databases that include transactional data collected at the point of sale. Research shows that access to complete individual level purchase histories is a prerequisite to successfully targeted marketing that encourages cross-buying, retains potential defectors, rewards high-value customers and increases referrals (Fader and Hardie 2009; Germann et al. 2014).

Yet practitioners and researchers often find that, whereas data accumulated under the "big data" paradigm comprise information about a growing number of

[1] U.S. Department of Defense News Briefing, February 12, 2002.

customers, the length of transaction histories is highly variable. Most individual purchase histories are relatively short or consist of long periods of low or no activity. This circumstance is especially present in noncontractual settings such as retail, which is characterized by constant fluctuations in the customer base. Fierce price competition, similar offerings among competitors, low switching barriers, aggressive promotional policies, and competing loyalty programs lead to customers readily and easily defecting to competing retailers (Rust, Zeithaml, and Lemon 2000, p. 102; Winer 2001). As a consequence, retail data often contain only relatively sparse purchase histories for the majority of customers. Processing this form of data with traditional marketing models is challenging, because it often results in specifications that feature more parameters than observations (Lenk and Orme 2009).

Moreover, in retail settings with no contractual relationships, a firm cannot even determine if a customer is just dormant or has already defected to a competitor. Thus, the reason for a defection often remains an "unknown-unknown" to the firm. Not only that, but abrupt changes in purchase behavior might also be due to "known-unknowns" that are hard for firms to foresee and could be rooted in the customers' personal lives such as relocation, change of income, or the start of a family.

These challenges shape both the models researchers have proposed and the methods practitioners apply to predict individual level purchase behavior. The academic models have continuously evolved over time from simple "recency, frequency and monetary value" heuristics (Roel 1988), the seminal Pareto/NBD model (Schmittlein, Morrison, and Colombo 1987) to stochastic models exploring the interrelation between transaction rates, spending, and customer drop-out (e.g., Glady, Lemmens, and Croux 2015; Mzoughia and Limam 2014; Schweidel and Knox 2013). Simultaneously, the development of Markov Chain Monte Carlo (MCMC) simulation techniques has given rise to hierarchical Bayesian models that allow to parsimoniously relate individual purchase behavior to cross-sectional heterogeneity and to obtain parameters estimates for complex models that previously have been difficult to cope with (e.g., Allenby, Leone, and Jen 1999; Aravindakshan, Rubel, and Rutz 2015; Guo 2009; Jen, Chou, and Allenby 2009; Jerath, Fader, and Hardie 2011; Schweidel, Park, and Jamal 2014). However, these models still lack the ability to address two major challenges that arise in noncontractual retail settings — (1) the inclusion of individual and overall seasonal effects and (2) the derivation of an accurate predictor of customer inactivity.

Many noncontractual settings exhibit strong seasonal and cyclical patterns of individual purchase behavior. What appears at the surface to be dormancy may be a naturally occurring pattern that will trigger purchasing when the next cycle comes around (Kumar and Reinartz 2012, p. 308). For example, if a customer only buys at a store when he or she is in town for vacation twice a year. Seasonality not only influences the accuracy of forecasting purchase levels, but also affects customer segmentation, individual targeting and the timing of marketing actions.

However, the models of the Pareto/NBD family (Fader and Hardie 2009) and related variants do no attempt to include individual and cross-sectional seasonal effects in the model framework. This situation is surprising, because Zitzlsperger, Robbert, and Roth (2009) find that forecasts of the standard Pareto/NBD can be improved if the results are post-hoc seasonally adjusted by a simple ratio to a moving average seasonal factor. Schweidel and Knox (2013) improve model fit by including a winter dummy-variable in their latent attrition model. And, while discussing the limitations of their evaluation of forecast algorithms, Ballings and Van den Poel (2015, p. 257) remark that "taking seasonality into account would only increase the predictive performance of our models."

Another challenge in noncontractual settings is to derive a probabilistic measure of a customer's unobservable latent state — the probability that the customer has defected and will not be active in the future. For this purpose Schmittlein, Morrison, and Colombo (1987) introduced the $P(alive)$ metric. Since then it has been customary for researchers to derive $P(alive)$ for models under the "buy 'til you die" paradigm (e.g., Fader, Hardie, and Shang 2010; Jerath, Fader, and Hardie 2011; Mzoughia and Limam 2014). This information about the customer's (unobservable) latent state has been used frequently, e.g., for optimizing customer reactivation campaigns (Ma, Tan, and Shu 2015), for benchmarking models (Batislam, Denizel, and Filiztekin 2007; Wübben and von Wangenheim 2008), for customer portfolio management (Sackmann, Kundisch, and Ruch 2010), for customer value analysis (Ho, Park, and Zhou 2006) or for examining the effect of modes of acquisition and retention on customer lifetime (Steffes, Murthi, and Rao 2008). Nonetheless, one can argue whether $P(alive)$ really provides meaningful managerial information as it assumes an infinite time horizon.

Practitioners often require forecasts for specific (finite) forecast horizons. As Wübben and von Wangenheim (2008, p. 91) state: "... it is of hardly any interest to managers whether a customer purchases after the planning horizon." Thus, current prediction models that rely on the $P(alive)$ metric do not provide information about the customers' inactivity in managerial relevant time frames and might be subject to systematic bias.

This thesis contributes to the literature stream of probability models for customer base analysis and follows an evolutionary model-building view (Fader and Hardie 2009). Specifically, I modify one existent and I develop two new models that address individual and overall seasonality effects and/or provide a way to accurately predict customer inactivity:

1. I put a twist on the classic Dynamic Model of Purchase Timing (DMPT), by Allenby, Leone, and Jen (1999) by proposing a refined method to more accurately predict inactive customers;[2]

2. I develop the Hierarchical Bayesian Seasonal Effects Model (HSM) that addresses the inclusion of seasonality effects. It is a hierarchical seasonal model that operates under the "always a share" assumption. The model relates individual purchase behavior to cross-sectional heterogeneity in purchase rates and seasonality;

3. Finally, I develop the Hierarchical Bayesian Seasonal Effects Model with Drop-Out (HSMDO) that addresses both the inclusion of seasonality effects and prediction of customer inactivity. The HSMDO features a joint seasonal purchase and customer lifetime model under the "buy 'til you die" assumption. It relates individual purchase behavior to cross-sectional het-

[2] This part of the thesis builds upon and extends my master's thesis "Prediction of Purchase Behavior: Benchmarking a Hierarchical Bayesian Model against Managerial Heuristics" (2010).

erogeneity in purchase rates, drop-out rates, and seasonality. In order to forecast customer inactivity with the HSMDO, I derive $P(Zero_F)$, the probability of observing no purchases for a specific customer in a given time frame F, as an alternative to $P(alive)$.

1.2 Goals of the Thesis

The goal of this thesis is to further the understanding of hierarchical Bayesian models by developing models that predict individual level purchase behavior in noncontractual settings. In order to simulate managerial decision making situations I follow the validation approach of Wübben and von Wangenheim (2008) and use the following managerial tasks for comparison: (1) forecasting short- and long-term purchase levels, (2) predicting which customers will be inactive in the future, and (3) rank ordering customers to identify the top 10% or 20% of the customer base.

I address the following key research questions regarding the DMPT and the newly proposed rule to predict customer inactivity:

1. Does the refined rule improve the accuracy of predicting inactive customers compared to the original rule and the hiatus heuristic?
2. How accurate is the DMPT in forecasting short- and long-term purchase levels compared to the baseline heuristic?
3. How accurate does the DMPT predict the future top 10% and 20% of the customer base compared to the baseline heuristic?

Regarding the HSM I address the following research questions:

1. How accurate does the HSM forecast short- and long-term purchase levels compared to SARIMA Models, the Holt-Winters Method (HW), and the baseline heuristic?
2. How accurate does the HSM predict the future top 10% and 20% of the customer base compared to SARIMA Models, the Holt-Winters method,

and the baseline heuristic?

3. Does the HSM provide useful information about overall and individual seasonality?

In regard to the HSMDO I address the following research questions:

1. How accurate does the HSMDO forecast short- and long-term purchase levels compared to SARIMA Models, the Holt-Winters Method, the baseline heuristic, the HSM, and the HMDO[3]?
2. How accurate does the HSMDO predict the future top 10% and 20% of the customer base compared to SARIMA Models, the Holt-Winters method, the baseline heuristic, the HSM, and the HMDO?
3. Does the HSMDO provide useful information about overall and individual seasonality?
4. Can a measure for the likelihood of customer inactivity in a finite time horizon be derived for the HSMDO?
5. Does $P(Zero_F)$ improve the out-of-sample classification accuracy of predicting inactive customers compared to $P(alive)$ and the hiatus heuristic?

The emphasis of the analysis is on individual customer forecasts, because they are essential for retail managers in deriving customer lifetime value and steering individual level marketing actions such as customer reactivation, customer reward, customer portfolio management, and cross-selling. Based on the newly developed models marketing managers will be able to accurately forecast customer specific future purchase patterns. They will be able to operate on a measure of individual seasonality that indicates how strongly the customer follows the cross-sectional seasonality, whether he purchases anti-seasonal or whether his behavior is non-seasonal. Such a measure is particularly useful for customer segmentation, targeting customers, and timing of marketing actions. Moreover, marketeers will be able to predict individual customer drop-out and might explore how likely each of their customers will buy in a definite time horizon of interest.

[3] The Hierarchical Model with Drop-out (HMDO) is a non-seasonal variant of the HSMDO. I use the HMDO to study the effects of different model parts on forecast accuracy.

1.3 Structure of the Thesis

The thesis is divided into nine chapters as depicted in Figure 1.1. The first chapter comprises the introduction including the scope and goal of this thesis as well as this section that lays out the structure of the dissertation.

In *Chapter 2*, I introduce relevant managerial concepts that shape the purpose and practical use of probabilistic models in customer relationship management. In particular, I discuss the link between loyalty and profitability, the relevance of customer value and, within this paradigm, core concepts such as customer lifetime value, customer equity, customer asset management as well as specific characteristics of relationships in noncontractual retail settings.

I outline the background of modeling and predicting purchase behavior in *Chapter 3*. I review the literature on modeling purchase behavior with a focus on probabilistic models for customer base analysis, which include individual purchase behavior and cross-sectional heterogeneity. In addition, I discuss marketing models that include seasonal components in the context of related time series concepts. Chapter 3 ends with a discussion of heuristic approaches for predicting purchase behavior and their relationship to simple stochastic models.

Chapter 4 provides an introduction to Bayesian theory and Markov Chain Monte Carlo methodology, because these methods are central to the model development and analysis. I discuss the various sampling algorithms, diagnostics for MCMC convergence, and measures of forecast accuracy.

Chapter 5 details the empirical settings, the three datasets I use for analysis, and data pre-processing. In addition to descriptive statistics, I depict the distributions of weekly interpurchase times and monthly purchase frequencies for each of the datasets. Also, exemplary customers purchase histories are shown to illustrate the individual level data on which the models operate.

Chapter 6 describes the Dynamic Model of Purchase Timing (DMPT) by Allenby, Leone and Jen 1999 and my improvement for deriving a probabilistic measure for customer inactivity. I recap the model framework consisting of (1) the hierarchical random-effects generalized gamma model, (2) the generalized gamma component

1	**Introduction** Scope, Goal, Structure of Thesis
2	**Conceptual Background on Customer Relationship Management** Scope and Goals, Customer Value, Noncontractual Relationships in Retail
3	**Background of Modeling and Predicting Purchase Behavior** Assumptions, Probability Models, Time Series Analysis, Heuristics
4	**Bayesian Modeling Approach and MCMC Methodology** Bayes‘ Theorem, Hierarchical Models, MCMC Simulation
5	**Research Data** Empirical Settings, Data Sets, Descriptive Statistics, Data Processing
6	**Dynamic Model of Purchase Timing** Model Framework, Refined Rule, Data Analysis, Results, Discussion
7	**Hierarchical Bayesian Seasonal Effects Model** Model Framework, Seasonal Effects, Data Analysis, Results, Discussion
8	**Hierarchical Bayesian Seasonal Effects Model with Drop-out** Model Framework, Drop-Out Process, Data Analysis, Results, Discussion
9	**Summary and Conclusions** Summary, Managerial Implications, Limitations, Future Research

Figure 1.1: Structure of this Thesis

mixture, (3) the temporal dynamics and link function, and (4) the priors for each parameter. I provide a mathematical formulation of my improved rule and perform parameter estimation and prediction. This chapter closes with a discussion of the results and compares them to managerial heuristics used in practice.

In *Chapter 7* I propose the hierarchical Bayesian seasonal model (HSM), which relates individual purchase behavior to cross-sectional heterogeneity in purchase rates and seasonality. I outline the new model framework, consisting of (1) the Poisson covariate gamma mixture, (2) hierarchical multiplicative seasonal effects, and (3) prior distributions and hyperparameters. I use Markov Chain Monte Carlo Gibbs sampling for estimation and prediction. Chapter 7 closes with a discussion of the new model's benchmark results.

The Hierarchical Bayesian Seasonal Model with Drop-out (HSMDO) is detailed in *Chapter 8*. I outline the model framework, consisting of (1) a periodic drop-out process with an SG/Beta mixture, (2) the Poisson covariate gamma mixture, (3) hierarchical multiplicative seasonal effects, and (4) prior distributions and hyperparameters. I provide a complete mathematical formulation of the model that includes the derivation of the individual level and sample likelihood function, predictive expectations of future purchase frequencies, and two measures of customer inactivity: $P(alive)$ and $P(Zero_F)$. I use Hamilton Monte Carlo (HMC) simulation with the no-U-turn sampler for parameter estimation and prediction. The chapter closes with a discussion of the new model's results.

The central findings are summarized in *Chapter 9*. I derive managerial implications for the use of probability models in customer relationship management. The chapter concludes with a discussion of the limitations of this thesis and an outlook towards future research on predicting customer behavior.

Chapter 2

Conceptual Background on Customer Relationship Management

2.1 Scope and Goals of Customer Relationship Management

Originally, the term Customer Relationship Management (CRM) was introduced in the 1990s within the field of Information Systems research (Payne and Frow 2005). It was used to describe technology-based solutions for managing the customer base with a tactical orientation such as salesforce automation (Law, Ennew, and Mitussis 2013; Payne and Frow 2005). The development of customer databases and communication technologies (Xie and Shugan 2001) has enabled firms to move beyond uniform marketing policies (Khan, Lewis, and Singh 2009) and realize the strategic potential in implementing customized customer base analysis tools (Boulding et al. 2005). Thus, Kumar and Reinartz (2012, p. 5) define customer relationship management as the "strategic process of selecting customers that a firm can most profitably serve and shaping interactions between a company and these customers."

The long-term goal of customer relationship management is to optimize the current and future value of customers by identifying valuable customers, increasing their loyalty, retaining them at the lowest possible cost, or reacquiring them if they left the relationship (Kumar, Bhagwat, and Zhang 2015). Thus, CRM implies the development of an understanding of customer behavior that can be used to refine and customize marketing instruments to increasingly fine segments or even to individual customers (Khan, Lewis, and Singh 2009; Peppers and Rogers 1993). For this purpose, CRM often includes the collection and maintenance of customer data (Ryals and Payne 2001) but does not necessarily require sophisticated analyses, concepts, or technology to analyze the customer base (Boulding et al. 2005).

Figure 2.1 displays the customer relationship management process according to Kumar and Reinartz (2012). Marketing managers start the process by analyzing the behavior of their active customer base. The analysis is shaped by the strategic goals of the firm and is the basis for developing a strategic marketing plan. This alignment between a firm's business strategy and the goals of the data analysis has been shown to be a prerequisite for a successful CRM strategy (Payne and Frow 2005). General business intelligence and concrete data analysis enable managers to derive a marketing plan that entails a number of individual marketing actions. For example, if managers find that their customers are especially prone to spend

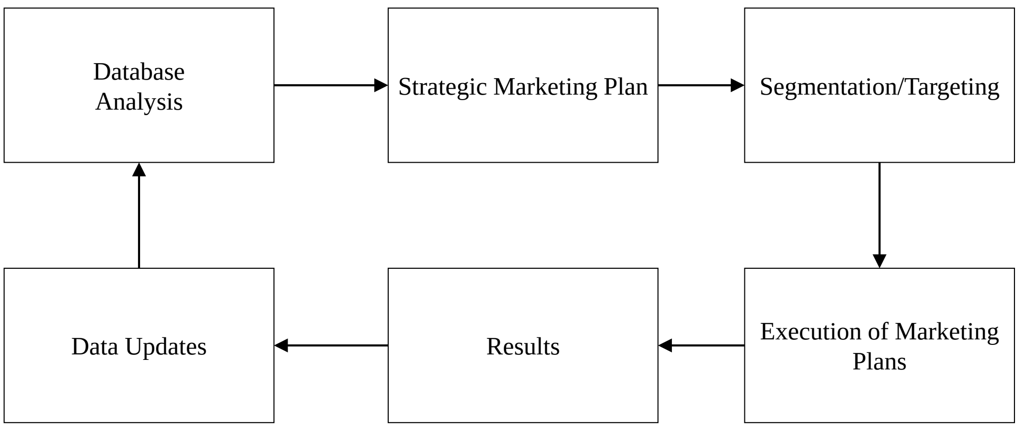

Figure 2.1: Layout of the Customer Relationship Management Process Own Illustration Based on Kumar and Reinartz (2012, p. 170)

more money in the weeks before winter holidays, they might develop a strategic plan to introduce a new product category in late November and to expand the new category's products sales over a period of years (Kumar and Reinartz 2012, p. 170).

Based on information stored in the customer database, managers are able to define a customer segment that is most likely to buy the items of the new product category. Moreover, they are able to select from the group of customers those who have spent above average in pre-Christmas seasons over the past years (Ballings and Van den Poel 2015). Subsequently, retailers can execute marketing programs by sending vouchers to selected customers or offer bundles to increase cross-buying. The next step in the CRM process is to monitor the success of those campaigns by analyzing changes in purchase behavior (Kumar and Reinartz 2012, p. 170) and to update the database. The whole process is then repeated permanently in order to continuously adapt the strategic marketing plan to changes in customer behavior, competitive environment, strategic goals, or shifts in supply (Neslin et al. 2013).

Research in marketing stresses the cross-functional importance of CRM for the firm, supports the positive impact of CRM practices on profitability and business performance (Boulding et al. 2005; Krasnikov, Jayachandran, and Kumar 2009; Ryals 2005) and showes that firms benefit from customer knowledge and improved customer satisfaction (Mithas, Krishnan, and Fornell 2005).

2.1.1 Customer Loyalty and Profitability

The goals of CRM are in line with the relationship marketing paradigm that emphasizes to foster customer loyalty in order to eventually increase the company's profits (Reichheld and Sasser Jr. 1990a). Customers show loyalty towards products, services, brands, or stores. Customer loyalty comprises behavioral loyalty such as repeat purchases (Brody and Cunningham 1968) as well as attitudinal loyalty that includes cognitive, affective, and conative elements (Oliver 1999). Attitudinal loyalty may manifest itself in a positive attitude towards a product, a high emotional attachment to a brand, or a high likelihood to recommend a service.

In the purchase decision-making process, a mostly cognitive problem-solving activity in which consumers move through a series of stages to make a purchase (Zinkhan 1992), attitudinal loyalty enables consumers to drastically reduce search time by omitting some alternatives and narrowing the focus to the brands, products or services to which they are loyal. Loyal customers are more likely to tolerate price increases and pay premium prices, to cross-buy and to spread positive word of mouth compared with short-term customers (Reichheld and Sasser 1991).

Attitudinal loyalty is related to customer satisfaction (Larivière et al. 2015). Both constructs focus on an overall ex post evaluation of a product, service or experience (Oliver 1997). Figure 2.2 illustrates the link between customer satisfaction, behavioral loyalty and profit as conceptualized by the service-profit chain (Heskett et al. 1994) and the satisfaction-profit chain (Anderson and Mittal 2000). Longitudinal research on the satisfaction-performance link (e.g., Bernhardt, Donthu, and Kennett 2000; Evanschitzky, von Wangenheim, and Wünderlich 2012; Gómez, McLaughlin, and Wittink 2004) supports the positive effect of customer satisfaction on operating profit over time.

The close relationship of satisfaction and loyalty and its impact on a firm's profits emphasizes the importance of customer retention (Datta, Foubert, and Heerde 2015; Reichheld, Markey Jr., and Hopton 2000). Customer retention is defined as a "customer's stated continuation of a business relationship with the firm." (Keiningham et al. 2007, p. 364). Customer retention management comprises a firm's efforts to continue it's relationships with customers. Already "committed" and,

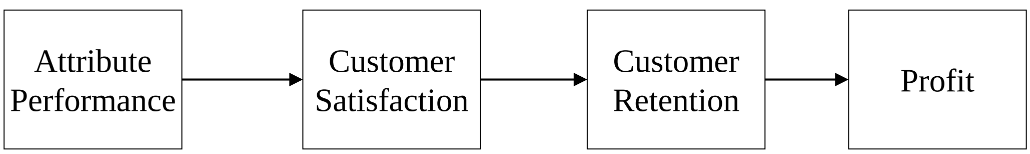

Figure 2.2: The Satisfaction-Profit Chain
Own Illustration Based on Anderson and Mittal (2000, p. 107)

thus, loyal customers can cost less to serve per period over their lifetimes than new customers, because a company does not incur acquisition costs (Kumar and Reinartz 2012, p. 28). Interest in customer retention and loyalty increased significantly after the work of Reichheld and Sasser Jr. (1990b) and Reichheld and Teal (1996), who found that a 5% increase in customer retention can increase a firm's profitability from 25% to 85%.

However, research has criticized this contention in regard to its generalizability, scope and implications (Dowling and Uncles 1997). Studies argue that the assumption of lower costs for maintaining existing customers compared with acquiring new customers only holds true for customers in a contractual relationship. In noncontractual relationships such as those of retail stores, frequent investments in customers (new and existing) are necessary, because customers may lose interest or switch easily to competitors (Dwyer 1997;Kumar and Reinartz 2012, p. 308).

Customer segments vary in their potential return on investment. Gupta and Zeithaml (2006) showed that 20% of customers may provide as much as 220% of the profits, implying that a large number of customers destroy value. This notion extends to specific customer segments that are loyal but still incur excessive resource allocation or exhibit high item return rates (Reinartz and Kumar 2002). Increasing retention or reward spending in these customers segments eventually leads to overspending and decreased profits (Haenlein, Kaplan, and Schoder 2006). Firms must make an effort to obtain information on individual or segment profitability. Therefore, an important part of CRM is measuring future customer value to identify profitable segments of customers and then developing specific strategies for reaching out and bonding to these customer groups.

2.1.2 Relationship Characteristics and Customer Retention

A framework that relates relationship characteristics to customer retention has been developed by Bolton, Lemon, and Verhoef (2004): the Customer Asset Management of Services (CUSAMS). The CUSAMS posits that it is useful to distinguish between length, breadth, and depth of customer relationships in order to derive valid predictors for explaining customers' value (Larivière and Van den Poel 2007).

The "breadth" dimension of customer relationships captures customers' "add-on" or cross-buying behavior, the number of additional and different services purchased from a company over time (Bolton, Lemon, and Verhoef 2004, p. 273). Larivière (2008) and Scherer, Wünderlich, and von Wangenheim (2015) show that the more different types of services a customer uses from a service provider, the less likely he or she is to defect.

The "depth" dimension is reflected in the frequency of service usage over time (Bolton, Lemon, and Verhoef 2004, p. 273). Bonfrer et al. (2007) find that the frequency of service usage positively affects customer retention probability in the cellular phone industry. The "length" of a relationship can be both consequence and driver of customer behavior (Bolton, Lemon, and Bramlett 2006; Reinartz and Kumar 2003; Zeithaml, Berry, and Parasuraman 1996). Lemon and von Wangenheim (2009) relate cross-buying to future usage of a service, showing that relationship breadth affects future relationship depth. Several studies show that the previous length of the customer relationship positively influences retention (e.g., Fader and Hardie 2007; Larivière 2008; Schweidel, Fader, and Bradlow 2008).

2.1.3 Metrics of Customer Value

The value the firm receives from implementing CRM measures is the increased success rate of applied acquisition, cross-selling and retention strategies that ultimately aim at heightening profitability of the firm (Payne and Frow 2005). Thus, firms are interested in tools to determine the most profitable and loyal customers as well as measures to estimate their future value to the firm. This need to determine the future long-term economic value of customers (Kumar and Reinartz 2012, p. 121) led to the rise of future-oriented metrics to quantify value, whereas traditional metrics currently used within business practice are often retrospective in nature (Zeithaml et al. 2006). Firms, for example, use measures of customer satisfaction based on customers' recent purchase experiences (Yi 1990), measures of service quality stemming from past encounters (Zeithaml 1999), or measures of attitudinal loyalty that reflect customer sentiment (Gupta and Zeithaml 2006). In general, such metrics provide managers with insights into why a firm is at its current state, but they do not fully capture and extrapolate ongoing and future developments such as customer defection and acquisition, market trends, direct marketing interventions, or seasonal effects (Esteban-Bravo, Vidal-Sanz, and Yildirim 2014). Retrospective measures offer only limited predictive ability in regard to future customer profitability (Petersen et al. 2009).

Forward-looking metrics of customer value address these issues and aim to guide company decisions with the goal of maximizing the long-term profitability of the customer base. The concept of Customer Lifetime Value (CLV) embodies this approach and constitutes a focal point of customer relationship management (e.g., Jain and Singh 2002; Mulhern 1999; Reichheld and Sasser Jr. 1990b; Rust, Lemon, and Zeithaml 2004). Customer Lifetime Value is the "the net present value of the profits linked to a specific customer once the customer has been acquired, after subtracting incremental costs associated with marketing, selling, production

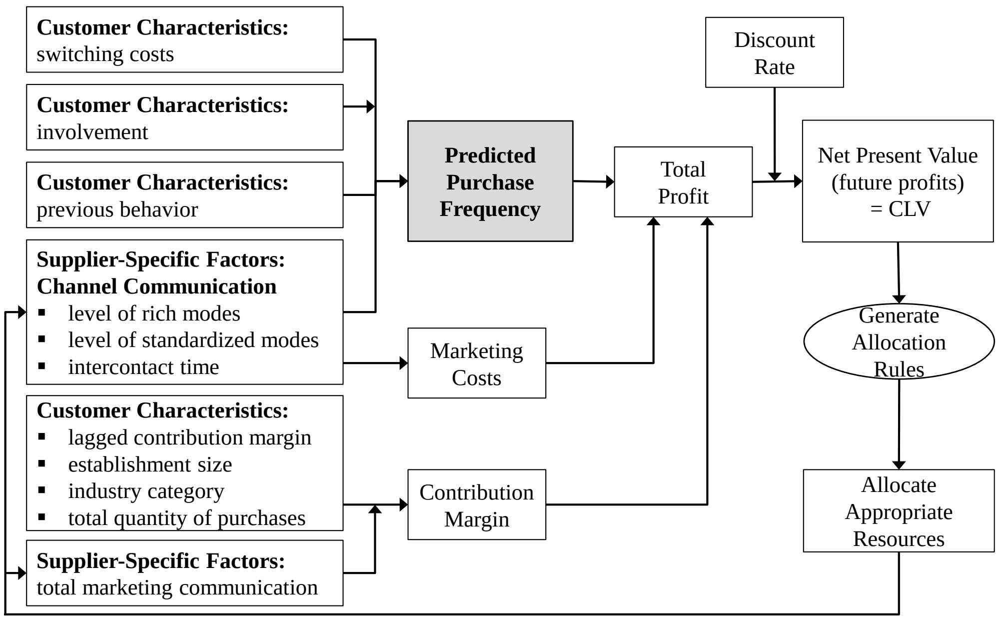

Figure 2.3: Conceptual Framework for Measuring CLV
Own Illustration Based on Venkatesan and Kumar (2004)

and servicing over the customer's lifetime." (Blattberg, Kim, and Neslin 2008, p.106).

Closely related to the CLV concept is Customer Equity (CE), which measures the future value of the total customer base. Rust, Lemon, and Zeithaml (2004, p. 110) define Customer Equity as "the total of the discounted lifetime values summed over all of the firm's current and potential customers." Thus, CE not only reflects forecasts of CLV for the current customer base but also includes forecasts of CLV for potential customers that are acquired in the future (Blattberg, Thomas, and Getz 2001; Blattberg and Deighton 1996).

Customers contribute to a firm's profitability not only by purchasing products and services but also by recommending the firm to others. Research documents the impact of word of mouth (WOM) on profitability and customer acquisition (Kumar, Petersen, and Leone 2010; von Wangenheim and Bayón 2007). A measure related to CLV that captures recommendation effects is Customer Referral Value (CRV). CRV measures "the net present value a customer creates for a firm via his or her referrals, i.e., the value a customer generates by referring the firm's products to its potential customers." (Kumar 2013, p. 23).

The ability to make individual level predictions of purchase behavior is essential for computing the future value of a firm's customers on a systematic basis (Fader and Hardie 2009). The CLV Framework (Figure 2.3) by Venkatesan and Kumar (2004) illustrates the central role of the prediction of purchase frequency in deriving CLV-based metrics. Thus, an accurate model for predicting individual-level purchase frequencies is a prerequisite for successful customer relationship management.

2.2 Customer Relationships in Retail Settings

2.2.1 Customer Buying Behavior in Retail

Studies identified different customer decision-making approaches in brick-and-mortar (Lysonski and Durvasula 2013) and online retail settings (Niu 2013). A common denominator among the identified individual decision-making approaches is that they fall along a continuum of various consumer characteristics or situational factors. These factors include the customer's degree of involvement, the perceived risk, the price of the products, the extent of effort necessary to obtain information about the products, and the customer's individual experience (e.g., Butler and Peppard 1998; Lamb, Hair, and McDaniel 2008).

Assael (1988) differentiates between four types of buying behavior: complex, variety-seeking, dissonance-reducing, and habitual. Sproles and Kendall (1986) identify eight mental decision-making styles based on a consumer styles inventory (CSI). The CSI inludes the styles of consumers who are driven by price, high quality, brand value, novelty and fashion, recreational shopping, impulse, or habit, as well as those, who are confused by choice. The novelty and fashion–conscious consumer matches the "variety seeker" type that appears to like new and innovative products and gains excitement from seeking out new things even while continuing to express satisfaction with previously purchased brands (Mowen 1988; Sproles and Kendall 1986). Bauer, Sauer, and Becker (2006) extended the CSI and differentiate an individual's decision-making style along different product categories and the respective product involvement.

A popular and parsimonious classification of individual decision-making styles distinguishes between extended, limited, and habitual decision making (Babin and Harris 2012, p. 252f; Lamb, Hair, and McDaniel 2008, p. 153; Solomon 2011, p. 335). The relation of habitual, limited and extended decision-making towards customer involvement, perceived risk, price, frequency, past experience, and information content is summarized in Figure 2.4.

Customers engage in extensive decision making when they are highly involved in

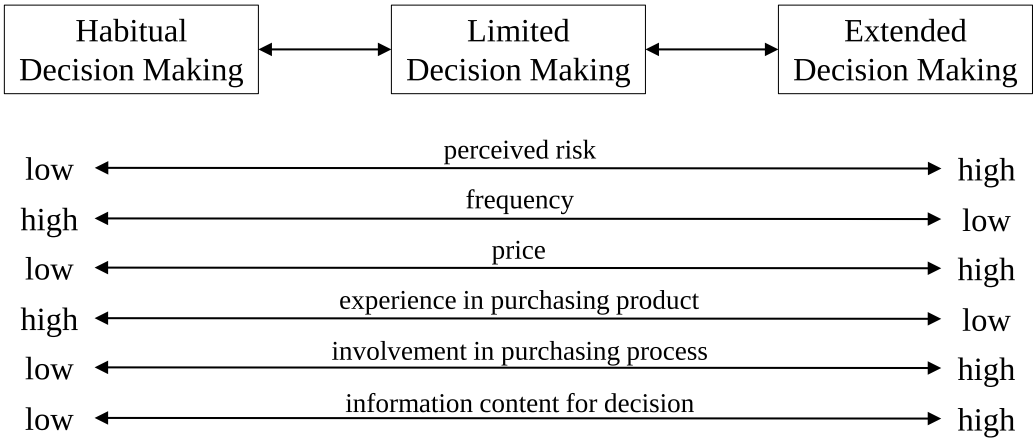

Figure 2.4: Habitual, Limited, and Extended Decision-Making
Own Illustration Based on Solomon (2011, p. 603)

a purchase and must undertake effort to search for information, evaluate alternatives and make a decision. This is often the case if the involved risk is high, such as when a customer buys a product for the first time or contemplates the purchase of an expensive item. Limited decision-making occurs when customers show less involvement and limit their search for information — for example, if customers are guided by prior beliefs about products and brands and restrict their evaluation to only a few alternatives based on very few attributes (Babin and Harris 2012, p. 252f).

In contrast, habitual decision making occurs when customers seek extremely little or no information at all. In this case, their purchase decision is based on habit to satisfy a recognized need. Habitual effects are related to consumers' recurring desires and often manifest themselves in a retail context as a customer's tendency to symptomatically re-purchase distinctive products or combinations of products (Shah, Kumar, and Kim 2014). Thus, such purchase patterns are the result of "behavioral persistence" (Reutterer et al. 2006).

2.2.2 Challenges of Noncontractual Relationships

There is a fundamental difference between CRM in contractual and noncontractual settings (Reinartz and Kumar 2000; Schmittlein, Morrison, and Colombo

1987). In a contractual retail setting such as food-box subscriptions or book-club memberships, a firm can easily observe whether a customer is still active or not. In this case, customers are usually contractually obligated to a minimum of regular purchases or a monthly fee. Customers, who decide to end their relationships with the firm, usually opt out of renewing their contracts or cancel them directly (Marinova and Singh 2014). These events enable the company to directly initiate marketing actions to regain such customers.

In noncontractual settings, however, the firm cannot determine if the customer is just dormant or has already defected to a competitor. Thus, the task to distinguish between active and potentially inactive customers is more complicated (Haenlein, Kaplan, and Beeser 2007). Moreover, in noncontractual settings such as retail the firm cannot expect a steady revenue stream from customers. Customers can show various patterns of purchase behavior at different times with long streaks of no or very little activity. Typically, retail is characterized by constant fluctuations in the customer base. Fierce price competition, similar offerings among competitors, low switching barriers, aggressive promotional policies, and competing loyalty

Opportunities for Transactions	Non-Contractual	Contractual
Continuous	Grocery Purchases Doctor Visits Hotel Stays Hairdresser Visits Subway Rides	Credit Card Student Mealplan Mobile Phone Usage Gym Membership
Discrete	Event Attendance Prescription Refills Charity Fund Drives Halloween Parade	Magazine Subs Insurance Policy Health Club M‘ship Cell Phone Plan

Type of Relationship With Customers

Figure 2.5: Classification of Customer Bases by Relationship Type
Own Illustration, Based on Fader and Hardie (2009)

programs lead to customers readily and easily defecting to competing retailers (Rust, Zeithaml, and Lemon 2000, p. 102; Winer 2001).

Fader and Hardie (2009) augment the distinction between contractual and noncontractual settings by taking into account the granularity of purchase opportunities (see Figure 2.5). The transactions within a contractual setting are not bound to the opportunities of contract renewal, e.g., usage-based billing for mobile phones or book clubs as described by Borle, Singh, and Jain (2008). If purchases can only occur at certain times, the opportunities for transactions are "discrete". For example, the renewal of a magazine subscription can usually be made only once a year. If purchases can occur at any time, such as in retail, or the renewal period is very short, the transactions are "continuous".

2.2.3 CRM Efforts in Noncontractual Settings

A CRM or marketing campaign typically comprises a series of interconnected promotional efforts. Typically, these campaigns are designed to market a new or existing product or service using a variety of marketing channels (Kumar and Reinartz 2012, p. 209). In this context, researchers recognize the value of customer analytics for targeting individual customers (Germann et al. 2014). However, in business practice, firms frequently analyze customer behavioral data only at an aggregate or customer segment level to initiate mass mailings or other CRM actions (Reutterer et al. 2006).

To facilitate strategic marketing campaigns directed at individual customers or groups of customers, a firm must determine how much it should invest in communication efforts (Jackson and Wang 1996, p. 45). For example, those individual customers the firm believes to have the highest future purchasing rates or the highest brand loyalty, would normally be selected first for retention programs (Shugan 2005; Winer 2001). Correspondingly, relationship managers should be prepared to abandon customers they believe to be unprofitable in the future (Haenlein, Kaplan, and Schoder 2006). In either case, an individual level prediction of future customer lifetime value is indispensable if a firm wants to assess the potential return of investment of its marketing campaigns.

The success of a marketing campaign involves reaching out to the right customer with the right offer at the right time and through the right channel (Kumar and Reinartz 2012, p. 210). The choice of communication channels is particularly important in targeted one-to-one marketing actions. Firms may choose one or more channels to communicate or interact: e.g., phone, direct mail, the Web, wireless devices, email, direct sales, or partner networks. Studies indicate that multichannel purchasing is positively associated with customer profitability and that properly designed marketing campaigns increase the number of multichannel customers (Montaguti, Neslin, and Valentini 2015).

Venkatesan and Kumar (2004) observe a non linear relationship between communication frequency and a customer's individual purchase frequency. They find that communication frequency beyond the optimal level results in diminishing returns. Thus, marketing managers must avoid "Type I" and "Type II" errors when executing marketing actions. Type I errors lead to lost revenue because the firm does not contact customers, who could have potentially provided revenue (Venkatesan, Kumar, and Bohling 2007). Type II errors are costly because the firm incurs costs for contacting customers who do not respond to the marketing instrument. In or-

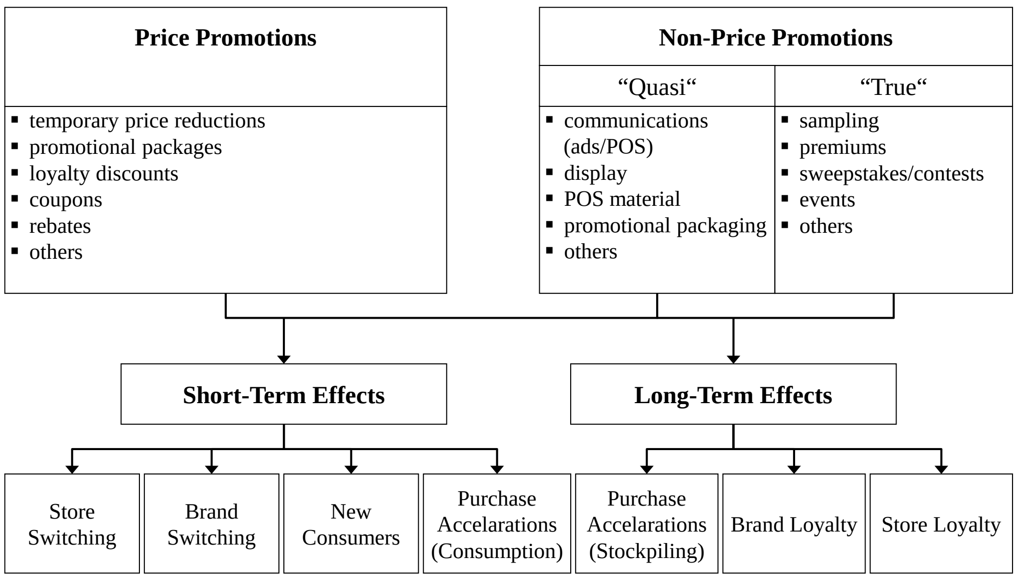

Figure 2.6: Short- and Long-Term Effects of Promotional Activities
Own Illustration Based on Gedenk (2002) [p. 19,104]

der to reduce Type II errors firms may customize the communication content to reach customers and get their attention, e.g., based on individual purchase history or customer demographic, lifestyle, and personal characteristics (Reutterer et al. 2006).

The choice of marketing instruments depends on the short- and long-term goals of the customer management efforts. Short-term goals may be to identify and acquire new customers, to induce customers to switch brands or stores, or to temporarily increase sales. Long-term goals of campaigns are to retain existing customers, to accelerate their long-term purchase behavior, or to increase their brand and store loyalty (Blattberg, Kim, and Neslin 2008, p. 128f; Gedenk 2002, p. 103f). Figure 2.6 gives an overview of different kinds of promotional efforts and their short- and long-term goals. Typically, CRM efforts focus on long-term effects such as strengthening relationships with customers and building loyalty (Breugelmans et al. 2015; Dorotic, Bijmolt, and Verhoef 2012; McCall and Voorhees 2010; Liu and Yang 2009).

Chapter 3

Background of Modeling and Predicting Purchase Behavior

3.1 Predicting Purchase Behavior Using Probability Models

The development of probability models to analyze customer bases serves two main purposes. First, the models help firms to understand customer purchase behavior and provide insights into the observed patterns of purchase. Second, firms can predict customers' future purchase patterns and customize marketing activities once they have a sound model of customer behavior.

The large number of factors that affect purchase behavior and the practical difficulties of capturing them accurately, simultaneously and in their totality give rise to a stochastic view of observed purchase patterns. The main assumption is that an observed purchase pattern is a superposition of behavior that is affected by a large number of preceding influence factors of which many are unknown. This makes the use of past transaction patterns for inference and prediction very attractive as purchase patterns are "widely available", tend to be "effective predictors" and can be "rich predictors" (Schmittlein and Peterson 1994).

An important feature of hierarchical models is that they relate individual level purchase behavior to cross-sectional purchase behavior. This theme is central to all models proposed in this thesis. In the following sections, I review the relevant literature, previous models and discuss the underlying assumptions. Also, I delineate the seasonal modeling approach taken in this thesis from previous marketing models and related time series literature.

3.1.1 "Always a Share" and "Buy 'Til You Die" Assumptions

Jackson (1985, p. 13) coined the terms "lost for good" and "always a share" to describe two types of relationships between an industrial vendor and its customers. Under the "lost for good" assumption, a customer is either totally committed to a vendor or completely lost and committed to another vendor. In contractual settings, firms often assume that a customer is "lost for good" when the customer does not renew a contract.

In contrast, the "always a share" model assumes customers to be permanently active, albeit at times at low purchase frequencies. The customer may have relationships with multiple vendors simultaneously, each vendor supplying a "share" of the customer's total need (Jackson 1985, p. 13).

In the realm of noncontractual retail settings, the "lost for good" model is often termed the "buy 'til you die" assumption (Fader and Hardie 2009; Schmittlein, Morrison, and Colombo 1987). Firms commonly assume that their relationship with a customer has only two states: "alive" and "dead". First, the customer is alive for an unobserved period of time ending with an unobserved "death" — for example, by defecting to another retailer. Only while he or she is "alive" the customer makes purchases; with death, the customer becomes permanently inactive and stops buying. The "lost for good" model classifies inactive or "dead" customers, who are won back, as new customers. A consequence of this approach is that individual metrics such as CLV are systematically understated for customers, who have reentered the customer base as "new" customers (Rust, Lemon, and Zeithaml 2004). Conversely, an "always a share" approach systematically overstates

CLV for customers, who have already defected, but the firm assumes that they will continue to purchase in the future.

Researchers model the "buy 'til you die" assumption with an individual lifetime or, alternatively, as an individual drop-out process. Such processes follow a purchase or interpurchase time model until the customer defects due to expired lifetime at the moment of the "death" event. Literature often uses the term "drop-out" synonymously with "death" in this context.

3.1.2 Pareto/NBD Family of Probability Models

The NBD or compound Poisson model assumes an "always a share" scenario (Ehrenberg 1959). Following this rationale, each customer is assumed to buy at a continued individual rate governed by a Poisson purchase process without ever defecting to another vendor.

The seminal Pareto/NBD model extended this reasoning with a "death" process under a "buy 'til you die" assumption. Here, the customer's individual lifetime follows an exponential distribution until he drops out and becomes permanently inactive (Schmittlein, Morrison, and Colombo 1987). The individual level purchase rates and the individual death rates follow different gamma distributions. Because the NBD model is nested inside the Pareto/NBD model, I present an adapted mathematical formulation based on Ehrenberg (1988, p. 128) and Schmittlein, Morrison, and Colombo (1987):

$$x_i \sim \quad Poisson(\lambda_i T) \quad \overset{\text{p.m.f.}}{\rightleftharpoons} e^{-\lambda_i T} \frac{(\lambda_i T)^{x_i}}{\Gamma(x_i+1)} \tag{3.1}$$

$$\lambda_i \sim \quad Gamma(r, \alpha) \quad \overset{\text{p.d.f.}}{\rightleftharpoons} e^{-\lambda_i \alpha} \lambda_i^{r-1} \frac{\alpha^r}{\Gamma(r)} \tag{3.2}$$

Each individual customer i purchases x_i items in time T. For each customer i, λ_i is his expected long-run purchase frequency per time unit. A gamma distribution of the λ_i's captures the heterogeneity across customers. The parameters of the group-level gamma distribution are the scale parameter $\alpha > 0$ and the shape parameter

$r > 0$. Integrating the λ_i's out of the gamma-Poisson mixture and setting $T = 1$ yields a negative binomial distribution (NBD) with two parameters r and α.[4] Hence the name NBD Model (Ehrenberg 1988, p. 128):

$$x \sim \quad NBD(r,\alpha) \overset{\text{p.m.f.}}{\rightleftharpoons} \int_0^\infty e^{-\lambda\alpha}\lambda^{r-1}\frac{\alpha^r}{\Gamma(r)}e^{-\lambda}\frac{\lambda^x}{\Gamma(x+1)}d\lambda \tag{3.3}$$

$$= \left(\frac{\alpha}{\alpha+1}\right)^r \left(\frac{1}{\alpha+1}\right)^x \frac{\Gamma(r+x)}{\Gamma(x+1)\Gamma(r)} \tag{3.4}$$

Schmittlein, Morrison, and Colombo (1987) extended the NBD model with a stochastic individual customer death process. The lifetime model assumes an exponential distributed lifespan $\tau_i > 0$ with an individual death rate μ_i for each customer. The death rates themselves are distributed gamma, with the group-level scale parameter $s > 0$ and the shape parameter $\beta > 0$:

$$\tau_i \sim \quad Exponential(\mu_i) \overset{\text{p.d.f.}}{\rightleftharpoons} \mu_i e^{-\mu_i \tau_i} \tag{3.5}$$

$$\mu_i \sim \quad Gamma(s,\beta) \overset{\text{p.d.f.}}{\rightleftharpoons} \frac{\beta^s}{\Gamma(s)}\mu_i^{s-1}e^{-\beta\mu_i} \tag{3.6}$$

The mixture of the exponential and gamma distribution yields, after integrating out the μ_i's, a Pareto II distribution of the customers' lifetimes (Schmittlein, Morrison, and Colombo 1987):

$$\tau \sim \quad Pareto(s,\beta) \overset{\text{p.d.f.}}{\rightleftharpoons} \int_0^\infty \frac{\mu e^{-\mu\tau}\beta^s\mu^{s-1}e^{-\beta\mu}}{\Gamma(s)}d\mu \tag{3.7}$$

$$= \frac{s}{\beta}(\beta/(\beta+r))^{s+1} \tag{3.8}$$

The combined purchase frequency and customer lifetime model by Schmittlein, Morrison, and Colombo (1987) reflects the "buy 'til you die" or, equivalently in this context, the "lost for good" paradigm. The NBD portion of the model describes the customers' expected purchase frequencies over their lifetime, while

4 Often the *NBD* is given as $x \sim NBD(r,p) \overset{\text{p.m.f.}}{\rightleftharpoons} (1-p)^r p^x \Gamma(r+x)/(\Gamma(x+1)\Gamma(r))$. The form used here is equivalent if one sets $\alpha = (1-p)/p$ with $0 < p \leq 1$.

the Pareto part of the model describes their exit from the relationship. Therefore, the complete model is called the Pareto/NBD model.

One interesting property of the model is that it allows to derive the probability that a customer i is still active or "alive" at the end of the observation period, conditional on the observed information (x_i, t_i, T). Here x_i denotes the number of purchases for customer i, t_i is the time of her last purchase, and T is the length of the observation period. Let now τ_i be the unobserved life time of customer i, then conditional on her individual purchase rate λ_i and drop-out rate μ_i the probability of being still alive at T is given by Schmittlein, Morrison, and Colombo (1987) as:

$$P(alive) = P(\tau_i > T | x_i, t_i, T, \lambda_i, \mu_i) = \frac{1}{1 + (\mu/(\lambda+\mu))(e^{(\lambda+\mu)(T-t)} - 1)} \tag{3.9}$$

Usually, the model parameters r, α, s and β, as well as the individual purchase and drop-out rates μ_i and λ_i, are estimated by maximum likelihood estimation. It should be noted that the expectation of the lifetimes τ diverges to infinity for $s \leq 1$ (Schmittlein, Morrison, and Colombo 1987):

$$E(\tau | s, \beta) = \int_0^\infty \tau \frac{s}{\beta} (\beta/(\beta+\tau))^{s+1} d\tau \tag{3.10}$$

$$= \begin{cases} \frac{\beta}{s-1} & \text{if } s > 1 \\ \infty & \text{if } s \leq 1 \end{cases} \tag{3.11}$$

The assumptions underlying the Pareto/NBD model are an active field of research. For example, Fader, Hardie, and Lee (2005a) replace the continuous-time Pareto drop-out process with a beta-geometric (BG) model. Drop-out opportunities are tied to each transaction, so that customer defection only occurs directly after a transaction with an individual probability of p_i. The cross-sectional heterogeneity over all p_i follows a beta distribution. The BG/NBD model simplifies the parameter estimation of the Pareto/NBD model while yielding similar results (Fader, Hardie, and Lee 2005a).

A variant of this model is the BG/GCP model by Mzoughia and Limam (2014). It keeps the BG customer lifetime model of the BG/NBD model, but the purchases during a customer's lifetime are assumed to follow a Conway-Maxwell-Poisson (COM-Poisson) distribution. This distribution has an additional parameter and allows to capture potential under and over dispersion of purchase data.

A further variation is the BG/BB model (Fader, Hardie, and Shang 2010), which, in addition to the above BG lifetime model, assumes transaction opportunities or reporting of transactions to occur at discrete points in time. This leads to a beta-Bernoulli (BB) distribution and, thus, the name BG/BB. This model improves prediction accuracy in cases where transactions of a company are associated with specific events or can only occur at fixed time intervals.

In their PDO/NBD model, Jerath, Fader, and Hardie (2011) generalize the Pareto/NBD with a discretized death process. At each point in discrete time, this model assumes a periodic death opportunity (PDO). In the limit, if the number of death opportunities goes to infinity, the model coincides with the Pareto/NBD. At the other extreme with no "death opportunity," the model collapses to an NBD-only model. Bemmaor and Glady (2012) propose a different continuous life-time process based on the Gompertz distribution with gamma mixing and find that the G/G/NBD Model provides a better fit to settings where offerings are strongly differentiated. Trinh et al. (2014) forgo a customer lifetime model and revisit the NBD assumption by developing a log-normal Poisson model for conditional trend analysis.

Researchers have applied the Pareto/NBD model and its derivatives for examining a wide variety of research questions in several different empirical settings. Van Oest and Knox (2011) modified the BG/NBD model by including complaint covariates to forecast the behavior of customers of a U.S. internet and catalog retailer that sells toys, novelties and party supplies. Abe (2009) examines a retailer that sells apparel, interior decoration, electronic toys, and gourmet food. He adapts the model by relaxing the independence assumption between purchase and the drop-out process.

Glady, Baesens and Croux's (2009) model includes a submodel for monetary value that they apply to retail bank datasets containing customers' stock exchange transactions. Batislam, Denizel, and Filiztekin (2007) proposed a zero-repeat-purchaser variant of the BG/NBD to analyze the customer base of a large grocery retailer in Turkey. Ho, Park, and Zhou (2006) augment the model with surveyed satisfaction to derive optimal investment in customer satisfaction measures. Fader, Hardie, and Lee (2005b) use the Pareto/NBD model for purchase frequency and an independent gamma/gamma submodel for transaction value to derive iso-value curves that identify customers with different purchase histories but similar CLVs. Researchers also use the Pareto/NBD to explore the impact of relationship characteristics and customer life-time on profitability (Reinartz and Kumar 2000; 2003).

The Pareto/NBD family of models and variants do not contain a hierarchical seasonal structure. As of yet no attempt to include individual and cross-sectional seasonal effects into the these model has been made. This situation is surprising, because Zitzlsperger, Robbert, and Roth (2009) find that the purchase frequency forecasts of the standard Pareto/NBD can be improved by post-hoc seasonal adjustment with a simple ratio to a moving average factor. Or, for example, Schweidel and Knox (2013) improve model fit by including a winter dummy-variable in their latent attrition model.

Moreover, it stands to reason that the drop-out process would also be highly susceptible to seasonal influences. If, for example, a Christmas gift shop observes at the end of December that the last purchase by one of his customers was made 3-month ago, then it is more likely that the customer has already dropped out than it would be if the firm had observed the same period of inactivity at the end of June.

3.1.3 Hierarchical Bayes: Generalized Gamma Models

The advent of the full hierarchical Bayesian approach and the Markov Chain Monte Carlo (MCMC) simulation technique not only freed researchers from many restrictive distributional assumptions but also had a large influence on model building and model complexity. The hierarchical structure of a fully specified Bayesian model parsimoniously mirrors the relationship between the individual level stochastic process and the cross-sectional heterogeneity. The MCMC methodology allows us to obtain parameter estimates in situations in which MLE is infeasible or very difficult.

Allenby, Leone and Jen's (1999) seminal "Dynamic Model of Purchase Timing" (DMPT) and its variants used the flexibility of the Bayesian framework to specify individual level interpurchase times as generalized gamma (GG) and the cross-sectional heterogeneity as inverse generalized gamma (IGG) distributed:

$$t_{ij} \sim \quad GG(\alpha, \lambda_i, \gamma) \quad \overset{\text{p.d.f.}}{\rightleftharpoons} \frac{\gamma}{\Gamma(\alpha)\lambda_i^{\alpha\gamma}} t_{ij}^{\alpha\gamma-1} e^{-(t_{ij}/\lambda_i)^{\gamma}} \tag{3.12}$$

$$\lambda_i \sim \quad IGG(\nu, \theta, \gamma) \quad \overset{\text{p.d.f.}}{\rightleftharpoons} \frac{\gamma}{\Gamma(\nu)\theta^{\nu\gamma}} \lambda_i^{-\nu\gamma-1} e^{-(1/\theta\lambda_i)^{\gamma}} \tag{3.13}$$

Here, t_{ij} is the j^{th} interpurchase time for customer i. The t_{ij} are assumed to be distributed generalized gamma with parameters $\alpha, \lambda_i, \gamma$. The individual customers' λ_i follow an inverse generalized gamma distribution (IGG) with parameters ν, θ, and γ to capture the cross-sectional heterogeneity. This part of the DMPT is a generalization of the Poisson-based NBD Model and the Erlang Model by Chatfield and Goodhardt (1973).[5]

Furthermore, the DMPT by Allenby, Leone, and Jen (1999) assumes a continuous mixture of GG/IGG components and temporal dynamics through a multinomial probit model that relates lagged interpurchase times to the mixture probabilities. Formulas 3.12 and 3.13 show only the one component case without the probit model. In Chapter 6, I discuss the full K-component model including the probit link and prior distributions.

The DMPT has been used in a variety of studies. For example, Kumar, Shah, and Venkatesan (2006) use the DMPT to examine the relationship between prior purchases and increased returns. Venkatesan and Kumar (2004) improve optimal customer selection and resource allocation in a CLV framework that uses the

[5] Poisson-distributed purchase frequencies are equivalent to interpurchase times that are exponentially distributed. The exponential distribution is a special case of the generalized gamma distribution with $\alpha = 1$ and $\gamma = 1$. The Erlang is a standard gamma distribution with an integer-value shape parameter.

DMPT to predict future purchases. Shu-Chuan (2008) identifies customer churn of an online karaoke service with a simplified DMPT aided by a recency heuristic. The model has been augmented by purchase quantity information to improve the accuracy of CLV-based customer selection (Venkatesan, Kumar, and Bohling 2007). Also, there has been some research activity to further develop the DMPT. For example, Guo (2009) proposes a variant of the DMPT that does not include a component mixture but instead adds product category as a covariate. Or, Jen, Chou, and Allenby (2009) propose a direct marketing model with temporal dependence of timing and quantity.

The DMPT features flexible distributional assumptions and temporal dynamics for the purchase process. Yet, this model does not feature an explicit drop-out process and, thus, does not directly provide information on a customer's likelihood of becoming inactive. A number of schemes have been proposed for the DMPT to use model inconsistent behavior as a proxy measure for customer drop-out (Reinartz, Thomas, and Kumar 2005; Wu and Chen 2000a; Allenby, Leone, and Jen 1999). I contribute to this literature by proposing a refined method to capture model inconsistent behavior, which aims at improving identification of inactive customers.

3.1.4 Alternative Models and Overview

The Pareto/NBD model family and Bayesian models related to the DMPT have inspired much research activity. Nevertheless, researchers have developed a wide variety of alternative approaches to model purchase behavior at an individual and group level. Some of these approaches employ different methodologies, use different assumptions or have different research foci.

For example, Mark et al. (2013) use a hidden Markov model to capture the evolution of customer dynamics in different customer segments over time. Romero, Van der Lans, and Wierenga (2013) propose a partial hidden Markov model that allows customer defection and reentry into the customer base. Meade and Islam (2010) use the Weibull distribution with gamma mixing and Plackett copulas to model dependence between interpurchase times. To model consumer responses to direct marketing Cui, Wong, and Lui (2006) develop a Bayesian network based on evolutionary programming. Gönül and Hofstede (2006) use hazard models to derive optimal catalog decisions with a log-lognormal baseline hazard while Manchanda et al. (2006) employ a piecewise exponential hazard function. Boatwright, Borle, and Kadane (2003) capture the purchase process through the Conway-Maxwell Poisson distribution, whereas Wu and Chen (2000b) developed an in-store-decision, 2-stage-logit model that accounts for nonpurchasers, zero repeaters (one-time buyers) and regular customers. In their model, the purchase behavior of regular customers follows an Erlang-k distribution.

Table 3.1 comprises an overview of probability models in the field of marketing relevant to this thesis. It includes the distributional assumptions of the purchase process and the model of cross-sectional heterogeneity. If a model features a drop-out process, the distribution of the life-time model and its cross-sectional distribution are denoted as well. I include for reference the classic Pareto/NBD model by Schmittlein, Morrison, and Colombo (1987) and the Dynamic Model of Interpurchase Timing by Allenby, Leone, and Jen (1999).

Table 3.1: Probability Models for Predicting Purchase Patterns and Behavior

Authors (Year)	*Title*	*Purchase P.*	*Heterogeneity*	*Drop-Out P.*	*Notes*
Glady, Lemmens, and Croux (2015)	Unveiling the relationship between the transaction timing, spending and dropout behavior of customers	Poisson	Gamma	Pareto	Gaussian copula between drop-out, spending and transaction rates
Mzoughia and Limam (2014)	An Improved BG/NBD Approach for Modeling Purchasing Behavior Using COM-Poisson Distribution	Conway-Maxwell-Poisson	Gamma	Beta-Geometric	Allows for over and under dispersion of purchase rates
Trinh et al. (2014)	Predicting Future Purchases with the Poisson Log-Normal Model	Poisson	Log-Normal	—	Conditional Trend Analysis Benchmark
Romero, Van der Lans, and Wierenga (2013)	A Partially Hidden Markov Model of Customer Dynamics for CLV Measurement	Poisson	Modified Gamma	Hidden Markov Model	Customers switch between activity, temporary inactivity and permanent defection
Mark et al. (2013)	Capturing the Evolution of Customer Firm Relationships: How Customers Become More (or Less) Valuable over Time	Poisson	Hidden Markov Model	Hurdle	Captures the evolution of different customer segments

(continued on the next page)

Authors (Year)	*Title*	*Purchase P.*	*Heterogeneity*	*Drop-Out P.*	*Notes*
Schweidel and Knox (2013)	Incorporating Direct Marketing Activity into Latent Attrition Model	Bernoulli	Geometric	Hidden Markov	Gaussian copula between interdonation times and donation amounts
Bemmaor and Glady (2012)	Modeling Purchasing Behavior with Sudden "Death": A Flexible Customer Lifetime Model	Poisson	Gamma	Gompertz Gamma	Life-time process with Gompertz distribution and gamma mixing
Jerath, Fader, and Hardie (2011)	New Perspectives on Customer "Death" Using a Generalization of the Pareto/NBD Model	Poisson	Gamma	PDO	Periodic death opportunities
Fader, Hardie, and Shang (2010)	Customer-Base Analysis in a Discrete-Time Noncontractual Setting	Bernoulli	Beta	Beta-Geometric	Transactions opportunities at discrete points in time
Meade and Islam (2010)	Using Copulas to Model Repeat Purchase Behaviour - An Exploratory Analysis via a Case Study	Weibull, Copula	Gamma	Multi-Stage	Copulas capture dependence between interpurchase times
Jen, Chou, and Allenby (2009)	The Importance of Modeling Temporal Dependence of Timing and Quantity in Direct Marketing	Bivariate Log-Normal	Gamma	—	Quantity as a dependent covariate

(continued on the next page)

Authors (Year)	*Title*	*Purchase P.*	*Heterogeneity*	*Drop-Out P.*	*Notes*
Abe (2009)	"Counting your Customers" One by One: A Hierarchical Bayes Extension to the Pareto/NBD Model	Poisson	Log-Normal	Exponential Log-Normal	Dependence of purchase and drop-out rates, adds covariates
Shu-Chuan (2008)	Online Customer Identification Based on Bayesian Model of Interpurchase Times and Recency	Gamma	Inverse Gen. Gamma	—	No temporal dynamics, uses recency for customer segmentation
Batislam, Denizel, and Filiztekin (2007)	Empirical Validation and Comparison of Models for Customer Base Analysis	Poisson	Gamma	Zero/Pareto	Adds a an additional chance for drop-out at time zero
Venkatesan, Kumar, and Bohling (2007)	Optimal Customer Relationship Management Using Bayesian Decision Theory: An Application for Customer Selection	Generalized Gamma	Inverse Gen. Gamma	—	Two lagged IPTs, joint timing-quantity model
Gönül and Hofstede (2006)	How to Compute Optimal Catalog Decisions	Log-Log Normal	Multivariate Normal	Proportional Hazard	Uses RFM and derives optimal decision rule
Cui, Wong, and Lui (2006)	Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming	Bayesian Network	Bayesian Network	Probability Networks	Network inputs are RFM and transaction variables
Manchanda et al. (2006)	The Effect of Banner Advertising on Internet Purchasing	Exponential	Multivariate Normal	Hazard	Covariates and exponential hazard

(continued on the next page)

Authors (Year)	*Title*	*Purchase P.*	*Heterogeneity*	*Drop-Out P.*	*Notes*
Fader, Hardie, and Lee (2005a)	Counting your Customers the Easy Way: An Alternative to the Pareto/NBD Model	Poisson	Gamma	Beta-Geometric	Defection only after transaction takes place
Boatwright, Borle, and Kadane (2003)	A Model of the Joint Distribution of Purchase Timing and Quantity	Conway-Maxwell-Poisson	Gamma	Memoryless Hazard	Relaxes independence of timing and quantity, two lagged time periods
Allenby, Leone, and Jen (1999)	A Dynamic Model of Purchase Timing with Application to Direct Marketing	Gen. Gamma	Inverse Gen. Gamma	—	Temporal dynamics, component mixture
Schmittlein, Morrison, and Colombo (1987)	Counting Your Customers: Who Are They and What Will They Do Next?	Poisson	Gamma	Pareto	Classic model of purchase patterns

Purchase P.: Purchase Process; Drop-Out P.: Drop-Out / Lifetime / Death Process

3.2 Forecasting Seasonal/Cyclical Time Series Data in Marketing

The field of time series analysis heavily influences marketing research on seasonal models. As of yet, time series forecast methods neither feature a seasonal hierarchical structure that accommodates hundreds or thousands of individual customers' purchase histories nor account for the possibility of customer drop-out. Specifically, the hierarchical approach of the Pareto/NBD or the DMPT that relates cross-sectional heterogeneity and individual level purchase behavior, as discussed in the previous Section, has not been extended to seasonality. With this thesis, I want to close this gap and propose new hierarchical seasonal models with and without customer drop-out to forecast purchase behavior. To explicate and delineate my seasonal models, I discuss marketing literature that considers seasonality in the context of time series concepts.

3.2.1 Simple Seasonal Covariates and Seasonal Adjustment

The most common methods for including seasonality in marketing models are the use of seasonal dummy variables as covariates (Wildt 1977) or prior seasonal adjustment of time series data (De Gooijer and Hyndman 2006). For example, Schweidel and Knox (2013) add a December dummy variable to a simultaneous latent attrition and direct marketing model to examine charitable donations to a nonprofit organization. Soysal and Krishnamurthi (2012) use a seasonal dummy for the Christmas season to analyze the demand of an apparel retail chain that supports seasonal fashion trends. The SCAN*PRO system, which is widely applied in retail to forecast brand unit sales, uses seasonal multipliers (Andrews et al. 2008).

The decomposition of a time series into seasonal and non-seasonal components allows the use of traditional forecast models on the adjusted non-seasonal component. This method is convenient for seasonal forecasting because the focal model can be used as is, without the need to add seasonal components. However, interpreting estimation results after seasonal adjustments requires some caution in comparison with the seasonal dummy approach.

The use of seasonal covariates and prior seasonal adjustment is related to the Frisch-Waugh-Lovell theorem (Lovell 1963). It states that a large class of seasonal adjustment procedures can be handled by seasonal covariates in a multiple regression framework. Since, the inclusion of seasonal covariates is preferable to the use of prior seasonal adjustment, because the latter entails an inherent tendency to overstate the significance of the regression coefficients without appropriate correction (Lovell 1963), the use of seasonal adjustment in marketing models has recently been sparse: Radas and Shugan (1998) seasonally adjust a time series by shortening and lengthening time units to model seasonal movie release patterns in the motion picture industry. Goodman and Moody (1970) use X11 seasonal adjustment to optimize a manufacturer's shipment quantities for price promotions with fixed durations. McLaughlin (1963) forecasts the sales of a manufacturing company and uses U.S. Bureau of the Census Method II to seasonally adjust the data. Semon (1958) forecasts sales of household textile goods through extrapolation of trend-adjusted average monthly differences.

The most widely used seasonal adjustment techniques are ratio-to-moving-average based and relate seasonal fluctuations proportionally to an overall or partial average. This approach also forms the foundation of the Method I, Method II X0-11, X11-ARIMA and X12-ARIMA methods as used by statistical agencies such as the U.S. Bureau of the Census (Findley et al. 1998).[6]

3.2.2 Exponential Smoothing: Holt-Winters Method

The use of exponentially weighted moving averages in time series analysis is attractive because of its minimal computational time and space requirements (Holt 2004b). For example, a forecast of purchase levels for the current period would only need to calculate a weighted average of two variables: the number of actual

6 Newer variants extrapolate a year of unadjusted data at each end of the series with an ARIMA model (see Section 3.2.3) —called "forecasting" and "backcasting"— in order to improve the accuracy of the moving average filter.

purchases during the previous period and the number of purchases forecasted for the previous period. Because the forecast for the previous period's number of purchases is in itself a weighted average, the resulting forecast is an exponentially weighted average of the complete series. This incremental update property and the declining weight put on older data[7] result in a simple, fast, and automatically adapting forecast scheme that smooths random fluctuations in a time series.

Holt (2004c)[8] extended the exponential averaging approach to decompose the series further into seasonal and trend components. Winters (1960) was the first to put Holt's theoretical work into practice by empirically testing the forecast accuracy of sales data of a construction company, a cooking utensil manufacturer and a paint wholesaler. Today, this approach of forecasting seasonal and trend components in time series analysis is called the Holt-Winters method (Holt 2004a;b; Ord 2004).

The Holt-Winters method and its derivatives are popular in a wide range of applications and have yielded robust results —for example, in the M3-competition (Makridakis and Hibon 2000). The method has been employed in marketing to optimize the use of marketing channels in a mining company (Strang 2012), to analyze the impact of item-level radio frequency identification (RFID) on stock-outs (Gaukler 2010), to forecast aggregate retail sales (Alon, Qi, and Sadowski 2001), and to forecast the number of tourists traveling to Hawaii (Geurts and Ibrahim 1975). In addition, Hardie, Johnson, and Fader (1993) use exponentially smoothed temporal reference prices to model loss aversion and preference dependence effects on brand choice.

The basic exponential model is described by the equations below, closely following Holt's (2004c) notation:

$$\bar{S_t} = A[S_t + (1-A)S_{t-1} + (1-A)^2 S_{t-2} + (1-A)^3 S_{t-3} ...] \tag{3.14}$$

$$\bar{S_t} = AS_t + (1-A)\bar{S}_{t-1} \tag{3.15}$$

[7] The weights are assumed to lay strictly between 0 and 1.

[8] The paper is a reprinted version of the 1957 report to the Office of Naval Research (ONR 52).

Here A is a constant between 0 and 1, S_t are the observed sales at time t and $\bar{S}_t$ is the estimate of the expected value of the distribution. The parameter A determines how fast the estimate adapts to changes. Equation 3.15 reflects the minimal amount of data needed to update the forecast. Only the expected value for the previous period $\bar{S}_{t-1}$ and the current level of sales are required to compute the new expected value for the current period $\bar{S}_t$.

$$\bar{S}_t = AP_tS_t + (1-A)\bar{S}_{t-1} \tag{3.16}$$

$$P_t = B\frac{\bar{S}_t}{S_t} + (1-B)P_{t-N} \tag{3.17}$$

The inclusion of ratio seasonals extends the basic model to equations 3.16 and 3.17 (Holt 2004c). $\bar{S}_t$ denotes the smoothed and seasonally adjusted sales rate. P_t is the seasonal adjustment ratio for the t^{th} periods. A and B are weights that determine how fast the level of sales (A) and the seasonal pattern (B) change. N is the length of the periodic pattern. The system of equations can be solved to yield the corresponding update formulas (Holt 2004c):

$$\bar{S}_t = \frac{A(1-B)}{1-AB}P_{t-N}S_t + \frac{1-A}{1-AB}\bar{S}_{t-1} \tag{3.18}$$

$$P_t = \frac{1-B}{1-AB}P_{t-N} + \frac{B(1-A)}{1-AB}\frac{\bar{S}_{t-1}}{S_t} \tag{3.19}$$

The Holt-Winters full model then adds a trend component to the seasonal model:

$$\bar{S}_t = AP_tS_t + (1-A)R_t\bar{S}_{t-1} \tag{3.20}$$

$$R_t = C\frac{\bar{S}_t}{\bar{S}_{t-1}} + (1-C)R_{t-1} \tag{3.21}$$

The trend component R_t is the adjustment ratio for the t^{th} period, and C is the parameter that determines the weight of past periods. The trend ratio implies a constant percentage change. Including the trend ratio yields the following update formulas:

$$\bar{S}_t = \frac{A(1-B)}{1-AB-(1-A)C} P_{t-N} S_t + \frac{(1-A)(1-C)}{1-AB-(1-A)C} \bar{S}_{t-1} R_{t-1} \tag{3.22}$$

$$P_t = \frac{(1-B)(1-(1-A)C)}{1-AB-(1-A)C} P_{t-N} + \frac{B(1-A)(1-C)}{1-AB-(1-A)C} \frac{\bar{S}_{t-1}}{S_t} R_{t-1} \tag{3.23}$$

$$R_t = \frac{AC(1-B)}{1-AB-(1-A)C} \frac{P_{t-N} S_t}{\bar{S}_{t-1}} + \frac{(1-C)(1-AB)}{1-AB-(1-A)C} R_{t-1} \tag{3.24}$$

Some authors prefer to use a non-lagged S_t in Equation (3.17). Ord (2004, p. 2) states that in order to produce a viable state-space model "... the use of lagged values should become standard practice, although the differences will be small."

Instead of the above multiplicative model, an additive variant can be derived in a similar fashion (Holt 2004c; Winters 1960). There have been numerous variations on the original Holt and Winters Model, most notably a dampened version by Taylor (2003). It addresses the observation that in long-term forecasting the persistence of the trend component leads to increasingly larger error (De Gooijer and Hyndman 2006).

Mentzer (1988) and Mentzer and Gomes (1994) propose an adaptive extended exponential smoothing that recalculates the model parameters after each forecast update. Simple exponential smoothing with drift is a variation of Holt's method with the trend parameter set to zero, whereas the "Theta" method is equivalent to exponential smoothing with the drift set to half the slope of the data's linear trend (De Gooijer and Hyndman 2006). Corberán-Vallet, Bermúdez, and Vercher (2011) formulated a seemingly unrelated regression model for multivariate time series, that includes the Holt-Winters model as a special case for forecasting industrial production in three communities in Spain.

A major drawback of these models in their original form, in spite of their solid forecast results, is the lack of uncertainty propagation. Hyndman et al. (2002) compiled a taxonomy of exponential smoothing models whose forecasts are equivalent to those of corresponding state-space models. The state-space models enable the computation of prediction intervals.

3.2.3 AR, ARMA, ARIMA and SARIMA Models

Linear stochastic models that describe the covariance structure of time series are widely used to analyze the dynamics of practically occurring phenomena. Specifically, the AR (autoregressive), ARMA (autoregressive moving average) and ARIMA (autoregressive integrated moving average) family of models has been employed in settings as diverse as telecommunication, traffic, employment, energy, truck sales, and health care tracking (De Gooijer and Hyndman 2006).

In marketing, Thaivanich, Chandy, and Tellis (2000) model the effects of direct television advertising for a toll-free referral service with an ARMA model. Franses (1991) forecasts the primary demand for beer in the Netherlands using an extended ARMA model. Holak and Tang (1990) use an ARIMA model to examine the influence of advertising on U.S. cigarette industry sales. This approach has been popularized by the works of Box and Jenkins, whose notation I will follow (Box and Jenkins 1976, p. 51):

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + ... + \phi_p \tilde{z}_{t-p} + a_t \tag{3.25}$$

Equation 3.25 is called an autoregressive (AR) process of order p. In this model, the current value of the process $\tilde{z}_t$ is a finite, linear combination of the previous values, the weights ϕ, and a shock a_t. The shocks are usually modeled as a white noise process with constant variance, mean zero and uncorrelated over time. The same relationship can be expressed using a backward shift operator B, which is defined as $B\tilde{z}_t = \tilde{z}_{t-1}$ and $B^m \tilde{z}_t = \tilde{z}_{t-m}$. Setting:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p \tag{3.26}$$

the AR process can be rewritten as:

$$\phi(B)\tilde{z}_t = a_t \tag{3.27}$$

Another perspective is to view the observed time series as dependent on a finite number of weighted shocks:

$$\tilde{z}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - ... - \theta_q a_{t-q} \tag{3.28}$$

This Equation (3.28) is called a moving average (MA) process of order q. If one analogously to $\phi(B)$ defines $\theta(B)$ in Equation (3.26), then one can write the MA model as:

$$\tilde{z}_t = \theta(B) a_t \tag{3.29}$$

In time series modeling it is often advantageous to combine the AR and MA models, because the combined model allows the representation of a time series with a lesser number of parameters. This approach leads to a mixed autoregressive-moving average model (ARMA) that can be written as:

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + ... + \phi_p \tilde{z}_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - ... - \theta_q a_{t-q} \tag{3.30}$$

$$\phi(B) \tilde{z}_t = \theta(B) a_t \tag{3.31}$$

In practice, adequate representation of actually occurring stationary time series can often be obtained with AR, MA or ARMA models with $p, q \leq 2$ (Greene 2008, p. 717; Box and Jenkins 1976, p. 11). Whereas finite MA processes are always stationary, the stationarity of the ARMA process depends on the AR part (Greene 2008, p. 721). The solutions of the characteristic Equation $\phi(B) = 0$ (treating B as a complex variable) are related to the stationarity and causality of the process. If all roots lie outside the unit circle, the process is stationary (Box and Jenkins 1976, p. 54). If some of the solutions are within the unit circle, the process may be stationary but not causal.

The ARMA framework can also be applied to a class of nonstationary processes if the d^{th} difference of the time series is stationary (Box and Jenkins 1976, p. 88):

$$\phi(B)(1-B)^d \tilde{z}_t = \theta(B) a_t \tag{3.32}$$

Equation 3.32 describes an autoregressive integrated moving average (ARIMA) process of order (p, d, q). The term "integrated" stems from interpreting the original time series as a summing up of the corresponding ARMA process d times (Box and Jenkins 1976, p. 55).

The principle of the d^{th} differencing operator $(1-B)^d = \nabla^d$ can be extended to seasonal differencing. If s is the seasonal lag (e.g., 12 for monthly data) and $\nabla_s = 1 - B^s$, then this equation relates a seasonal component to past seasonal components (e.g. a month to the same month in past years) (Box and Jenkins 1976, p. 304):

$$\Phi(B^s)_Q \nabla_s^D \tilde{z}_t = \Theta(B^s)_P \alpha_t \tag{3.33}$$

The errors α_t are now correlated (e.g., consecutive months of one particular year) and can be modeled with the standard ARIMA process:

$$\phi(B)_p \nabla^d \alpha_t = \theta(B)_q \alpha_t \tag{3.34}$$

Box and Jenkins (1976) call Equation (3.33) the "between periods" development of the series and Equation (3.34) the "within periods" part of the process (p. 323). The resulting multiplicative process (SARIMA) is of order $(p,d,q) \times (P,D,Q)$.

The ARIMA family of models has been extended to multivariate time series data in the form of vector AR (VAR), vector ARMA (VARMA) and vector ARIMA (VARIMA) models. De Gooijer and Hyndman (2006) note that in general VAR models tend to suffer from "overfitting" and can provide poor out-of-sample forecasts, even though within-sample fit is good. One method of addressing these issues is to put Bayesian priors on the parameters and their structure (BVAR). Gefang (2014) uses a Bayesian doubly adaptive elastic net lasso for VAR shrinkage. Canova (1993) proposes a BVAR for series that possess common patterns at seasonal or other frequencies to forecast quarterly data of industrial production in three European countries.

Unfortunately, these multivariate extensions typically assume that the individual time series are dependent on each other and provide reliable estimates only for a small number of interrelated time series. For example, Moriarty and Salamon (1980) forecasts sales data of branded goods in four states by relating four ARMA models through a seemingly unrelated regression. Takada and Bass (1998) performed multiple time series analysis with a VARMA model to examine three firms and their competitive marketing behavior in an oligopolistic market. In contrast, the Pareto/NBD and DMPT model families are used in marketing to forecast individual purchase behavior and cross-sectional heterogeneity of hundreds or thousands of customers.

3.2.4 Error-Correction, State-Space and Alternative Models

Researchers have used a number of other approaches for modeling time series data that allow the inclusion of seasonal effects. Engle and Granger (1987) introduced error-correction (EC) models for cointegrated time series analysis that relax some of the assumptions of VARIMA models. For example, Franses (1994) extends a deterministic Gompertz process with covariates and finds an error-correcting form to model Dutch new car sales.

Another approach consists of state-space models characterized by relating system and measurement equations to estimate state variables of dynamic systems (Xie et al. 1997). The system equation is alternatively called a transition equation because it describes the evolution of the state over time (Dagum and Quenneville 1993). The most well-known algorithm for these models is the Kalman filter, which provides efficient one-step-ahead forecasts, prediction errors and associated variances (De Gooijer and Hyndman 2006).

Van Everdingen, Aghina, and Fok (2005) use an augmented Kalman filter for examining cellular phone adoption in 15 EU countries. Xie et al. (1997) use this state-space approach to model the diffusion of new products for consumer durables, medical equipment, and educational programs. Dagum and Quenneville (1993) propose a seasonal Kalman filter based state-space model and apply it to department store sales in Canada. Raynauld and Simonato (1993) propose a seasonal Bayesian VAR (BVAR) using the Kalman filter framework to model the coevolution of eight economic variables of the Federal Reserve Bank of Minneapolis. One advantage of the state-space approach is that it provides a unifying framework that can express any linear model, as such many of the exponential smoothing models can be restated as state-space models (De Gooijer and Hyndman 2006).

Dynamic nonlinear or regime-switching models represent another form of state-space models —– ones that do not use the Kalman filter. Park and Gupta (2011) use a state-space regime-switching model to analyze yoghurt sales and find that price promotions offered during periods with a high purchase likelihood result in greater sales increases than promotions offered during a period with low purchase likelihood.

3.2.5 Hierarchical Time Series Analysis and Cross-Sectional Heterogeneity

A hierarchical time series consists of multiple individual time series that are hierarchically organized and can be aggregated at different levels (Hyndman et al. 2011). For example, all product sales of a brand in one store aggregate to the brand-level sales for that store. All brands aggregate to store-level overall sales, and all store-level sales in one country aggregate to national sales.

Typical schemes for hierarchical time series analysis operate either bottom-up or top-down. The bottom-up approach forecasts the lowest level data first and then aggregates these lower-level forecasts to higher levels. With the top-down approach, an aggregate forecast at the highest level is disaggregated to the lower levels. Whereas there is little research that includes seasonal effects for these models, a notable exception is Withycombe (1989), who used a top-down approach to forecast sales demand of peripheral equipment for computer systems. He first aggregates all products to an overall product line sales history, then derives the seasonal indices from these aggregated data and finally propagates this information top-down, back to the individual products. Hyndman et al. (2011) propose a combined approach that forecasts on all levels and then combines and reconciles these forecasts.

The modeling of cross-sectional heterogeneity, which is a main feature of the Pareto/NBD and DMPT-related marketing models of individual purchase behavior (see Section 3.1.2), can also be interpreted as a combined bottom-up, top-down approach. For example, the NBD model views the purchase process on an individual customer level as a Poisson process and captures the cross-sectional heterogeneity by a higher-level gamma distribution on the Poisson parameters. A related examples from time series analysis literature is Böckenholt (1998), who proposes a finite mixture of integer-valued autoregressive Poisson models to analyze scanner panel data of powder detergents. His model facilitates the analysis of heterogeneity and serial correlation with the effects of covariates. However, researchers have yet to extend these approaches to include seasonal effects.

3.3 Heuristic Approaches for Predicting Customer Behavior

3.3.1 Managerial Perspective on Heuristics

Shah and Oppenheimer (2008, p. 207) refer to heuristics as "methods that use principles of effort-reduction and simplification", whereas Gigerenzer and Gaissmaier (2011, p. 454) see heuristics as "a strategy that ignores part of the information, with the goal of making decisions more quickly, frugally, and/or accurately than more complex methods." Heuristics can also be characterized as simple and plausible psychological mechanisms of inference that a mind can carry out under limited time and knowledge (Gigerenzer and Goldstein 1996) or simply as "Rules of Thumb" (Fox 2015, p. 82)

These properties make heuristics particularly attractive tools for business managers, who find themselves constrained by limited time and knowledge (Newell, Weston, and Shanks 2003). In increasingly dynamic market environments, where strategies that were successful in the past may no longer work in the future, heuristics provide fast, frugal and robust methods of inference (Lee and Cummins 2004), especially when considering the opportunity costs managers face (Rieskamp and Hoffrage 2008). Heuristics exploit recurrent features of the environment, while saving costs of information search and integration (Goldstein and

Gigerenzer 2009; Gigerenzer 2006). The notion of an cognitive "adaptive toolbox", a repertoire of fast and frugal heuristics, is gaining traction in academia and practice (Gigerenzer 2001; Gneezy, Gneezy, and Lauga 2014; Sevilla and Kahn 2014).

There is some evidence from cognitive psychology that heuristics can perform surprisingly well under external constraints (Rieskamp and Hoffrage 2008). For some tasks heuristics outperform complex strategies that require more information and computation (Gigerenzer and Gaissmaier 2011). In area of marketing Huang (2012) finds that heuristics can perform at least as well as a probabilistic model for customer prioritization; Wübben and von Wangenheim (2008) could not support superior performance of the Pareto/NBD compared to simple heuristics in a number of scenarios; and Hagerty (1987) shows that under certain conditions naive models' forecasts can be as good as those of more complex linear models. It is noteworthy that Marshall (2015) proposes a heuristic to approximate the parameters of the Pareto/NBD Model.

Models and heuristics are related by the fact that every heuristic can be stated as a model, albeit a simple one. Makridakis and Wheelwright (1977) note that it is not uncommon for naive forecast models to provide adequate accuracy in certain situations and that more sophisticated methods may not give sufficient improvement in relation to cost. They also emphasizes the usefulness of naive models as a basis for comparing alternative approaches. For this reason, I include a number of commonly used heuristics as a benchmark in my thesis. These heuristics operate in the realm of the recency, frequency and monetary-value (RFM) framework.

3.3.2 Recency, Frequency and Monetary Value Heuristics

The classic RFM approach characterizes customer behavior and customer value based on information about individual customers' past purchase behavior (Kumar and Reinartz 2012, p.111). RFM based heuristics typically do not rely on complete and detailed purchase histories, but on aggregate measures that reflect how recently and how frequently a customer purchases, and on the total amount of his spending (Colombo and Jiang 1999).

RFM stands for the following three metrics upon which customers are evaluated:

1. Recency is a measure of how much time has elapsed since a customer's last purchase;
2. Frequency is the number of purchases by a customer in a particular time frame; and
3. Monetary value stands for the average amount spent by the customer in past transactions.

There are many widely applied managerial heuristics that operate on one or more of these three metrics under the general assumption that customers will continue to behave according to their previous behavior (Venkatesan and Kumar 2004). For example, assuming that a customer will continue to purchase at his past average purchase rate is a basic heuristic.

Companies often rank-order their customers according to purchase frequency or the average monetary value of purchases to prioritize their customer investment (Roberts and Berger 1999). Such a heuristic is based on the believe that currently high-ranked customers account for a substantial part of a company's future profit and that low-ranked customers are believed to be responsible for a large share of a company's future costs (Mulhern 1999). This practice is closely related to the 80/20 heuristic that expresses the common belief that 80% of a firm's profits come from the top 20% of the customers. A more recent variant is the 220/20 rule. It states that 20% of the customers provide 220% of the profits, implying that a large number of customers destroy value (Gupta and Zeithaml 2006).

A rule of thumb that operates solely on recency is the "hiatus" heuristic, which predicts if a customer is about to defect (Gigerenzer and Gaissmaier 2011). Hiatus describes the amount of time that has passed since a customer's last purchase. If this timespan exceeds a certain cutoff value, a company considers a customer to be at risk of defection or as already defected. Airlines, apparel retailers and industrial buyers use this heuristic to initiate targeted marketing actions (Wade 1988; Wübben and von Wangenheim 2008; Schmittlein and Peterson 1994).

Combined RFM approaches that take all three metrics into consideration leave the realm of simple heuristics, but have gained some attention from practitioners and researchers. A common procedure described by Gönül and Hofstede (2006) is to assign customers a score on each of the RFM dimensions. Then customers are grouped based on their respective RFM scores. For example, customers are divided into five groups of equal size according to the recency metric. The topmost group is assigned a recency score of 1, the next group is assigned a score of 2 and so forth, until the bottommost group is assigned a score of 5. The same grouping and coding is then performed on frequency and monetary value metrics. In the end, there will be $5 \times 5 \times 5$ RFM score-groups to which a customer may belong.

The RFM technique is helpful to organizations for identifying and targeting valuable and for avoiding investments in unprofitable customers. It has seen widespread industry applications (Berry and Linoff 2004; Hughes 1996; Reutterer et al. 2006). For example, mail-order businesses often use RFM metrics to determine which customers should get a catalog (Gönül and Hofstede 2006).

To improve segmentation results, researchers extended the RFM technique with additional behavioral measures (Kao et al. 2011), RFM pattern recognition (Hu and Yeh 2014), weighing of RFM variables (Nikumanesh and Albadvi 2014), or more recently the "clumpiness" of purchases (Zhang, Bradlow, and Small 2015; Kumar et al. 2015).

Chapter 4

Bayesian Modeling Approach and MCMC Methodology

4.1 Hierarchical Models and Bayesian Inference

The challenge in fitting a hierarchical model is estimating the data-level coefficients along with the higher-level model parameters (Gelman and Hill 2007, p. 345). Hierarchical models can become very complex and traditional asymptotic methods provide solutions only to specific problems without generalization (Ntzoufras 2009, p. 1).

The most direct and parsimonious way to approach hierarchical models is through "Bayesian inference", a method where higher-level model parameters are treated as prior information in estimating lower-level coefficients (Gelman and Hill 2007, p. 346). For a long time the intractabilities involved in calculation the posterior distribution limited Bayesian inference to simple models with mostly conjugate priors that could be handled analytically (Rossi, Allenby, and McCulloch 2005, p. 1). The development of Markov Chain Monte Carlo (MCMC) methods changed this, it is now possible to estimate complicated models that describe and solve problems that could not be solved before through traditional means (Gelfand and Smith 1990).

The Bayesian approach considers all parameters θ of a model as random variables with prior distribution $\pi(\theta)$ (Ntzoufras 2009, p. 1). In contrast, traditional statistical inference regards unknown parameters as fixed constants and infers an estimate, given the data D, by maximizing the likelihood $P(D|\theta)$.

Prior distributions in Bayesian statistics can be interpreted in various ways, for example as results of previous experiments, subjective knowledge or as quantified ignorance in form of a non-informative or vague prior. The prior distribution $\pi(\theta)$ is combined with the traditional likelihood $P(D|\theta)$, given the data D, to obtain the posterior distribution $\pi(\theta|D)$ of the parameter of interest θ by application of Bayes' Theorem:

$$\pi(\theta|D) = \frac{P(D|\theta)\pi(\theta)}{P(D)} \tag{4.1}$$

$$P(D) = \int_\theta P(D,\theta)d\theta = \int_\theta \pi(\theta)P(D|\theta)d\theta \tag{4.2}$$

The equation can be restated in proportional terms, because $P(D)$ is just a normalizing constant that keeps the integral of the posterior density equal to one and contains no information in itself:

$$\pi(\theta|D) \propto P(D|\theta)\pi(\theta) \tag{4.3}$$

The Equation reflects the Bayesian learning process: Posterior knowledge $\pi(\theta|D)$ about a parameter θ is the result of prior knowledge $\pi(\theta)$ and new data D in view of the prior $P(D|\theta)$.

The posterior $\pi(\theta|D)$ only depends on the data through the likelihood $P(D|\theta)$ implying that Bayesian inference adheres to the likelihood principle (Gelman et al. 2003, p. 9). The likelihood principle posits that two likelihood functions for a given sample contain the same information about θ if they are proportional to one another (Berger and Wolpert 1988, p. 19).

It is possible to obtain classical point estimates through the use of large sample asymptotics. If the sample size n of the data D, for example through simple replication, increases to infinity the posterior $\pi(\theta|D)$ approaches normality with mean θ_0 and variance $(nJ(\theta_0))^{-1}$ under certain regularity and identification conditions

(Gelman et al. 2003, p. 107). J denotes the Fisher information matrix. Because $n \to \infty$ additionally implies that the likelihood dominates the prior and that the distance between the posterior mode and θ_0 approaches zero, one obtains an estimate for the maximum likelihood (Lunn et al. 2012, p. 54). This result, which is also known as the Bernstein van Mises Theorem or Bayesian Central Limit Theorem, allows the Bayesian approach to produce classical ML estimates and confidence regions if desired (Boucheron and Gassiat 2009).

The Bayesian approach is particularly well suited for complex hierarchical models, because higher-level model parameters can act as priors for lower-level parts of the model. In a Bayesian framework, this concept can be applied recursively over a number of levels, down to the data-level model. The parameters defined on the highest level of the model are called hyper-parameters and specify the initial, often non-informative, prior distributions. The ability to parsimoniously specify and estimate nested hierarchical models of high complexity is the reason why such models are often referred to as hierarchical Bayesian (HB) models (Allenby, Bakken, and Rossi 2004).

Purchase frequency models, for example, assume that observed individual purchases reflect an underlying process that, if understood, allows the prediction of future individual customer behavior. At the same time, on a higher-level, these individual processes may be linked through a cross-sectional model of all customer's behavior. The NBD model (Ehrenberg 1959), for example, on the data-level, assumes that observed purchases x_i in time T by an individual customer i are distributed Poisson with purchase rate $\lambda_i T$:

$$x_i \sim \quad Poisson(\lambda_i T) \ \stackrel{\text{p.m.f.}}{\rightleftharpoons} e^{-\lambda_i T} \frac{(\lambda_i T)^{x_i}}{\Gamma(x_i + 1)} \tag{4.4}$$

On the next higher level, the model assumes customers to be heterogeneous in their purchase rates according to a gamma distribution. In a Bayesian framework the cross-sectional heterogeneity of the NBD model can be seen as a gamma prior on the λ_i's:

$$\lambda_i \sim \quad Gamma(r, \alpha) \ \stackrel{\text{p.d.f.}}{\rightleftharpoons} e^{-\lambda_i \alpha} \lambda_i^{r-1} \frac{\alpha^r}{\Gamma(r)} \tag{4.5}$$

A third level, that specifies the priors and their hyper parameters for the lower level gamma distributions would complete a Bayesian formulation of the model as for example the hierarchical Bayesian version of NBD model by Lichung, Chien-Heng, and Allenby (2003).

4.2 Markov Chain Monte Carlo Simulation

The marginal posterior densities of the parameters of interest are difficult or often impossible to derive analytically for even simple hierarchical models of practical relevance (Ntzoufras 2009, p. 32). Additionally, hierarchical marketing models that feature several parameters for each individual customer span a high-dimensional parameter space with at least as many dimensions as there are customers (Rossi, Allenby, and McCulloch 2005, p. 49). Fortunately however, iterative approximation through MCMC simulation provides by far the most powerful and flexible class of algorithm to yield accurate estimates of the posterior distribution even for complex multi-level models with high-dimensional parameter spaces (Lunn et al. 2012, p. 58).

The main idea of the MCMC approach is to reduce the problem of sampling the parameter vector θ from a high-dimensional posterior distribution of unknown and often intractable form, to sequentially sampling from simpler univariate or low-dimensional distributions. The goal is to approximates the posterior distribution $\pi(\theta|D)$, the target distribution, by the limiting invariant distribution of the Markov sequence (Gelman et al. 2003, p. 286).

A sequence of random variables $\{\theta^{(t)}\}_{t\geq 0}$ is called a Markov Chain, if:

$$P(\theta^{(t+1)}|\theta^{(t)},\{\theta^{(n)}:0\leq n\leq t-1\})=P(\theta^{(t+1)}|\theta^{(t)}) \tag{4.6}$$

This states that for any $\theta^{(t)}$, the past $\{\theta^{(n)}:n\leq t-1\}$ and the future $\{\theta^{(n)}:n\geq t+1\}$ are independent, and that the present $\theta^{(t)}$ only depends on $\theta^{(t-1)}$ (Neal 1993, p. 36). A proper Markov Chain is defined by the initial $\theta^{(0)}$ and by the transition kernel $P(\theta^{(t+1)}|\theta^{(t)})$ that specifies the conditional distribution of $\theta^{(t+1)}$

given $\theta^{(t)}$. For repeated transitions, if $t \to \infty$, a large class of Markov Chains are ergodic and converge to their equilibrium or invariant distribution under mild assumptions of irreducibility and aperiodicity (Neal 1993, p. 38, 45). Under these conditions, the independence of future and past θ additionally ensures that the resulting target distribution is independent of the initial $\theta^{(0)}$ (Lunn et al. 2012, p. 63). A number of sampling algorithms have been devised to produce such irreducible and aperiodic Markov Chains.

The Gibbs sampler by Geman and Geman (1984) and Gelfand and Smith (1990) is an algorithm for generating an irreducible, aperiodic and Markov Chain that has a high dimensional stationary distribution, that only requires sampling from univariate or low-dimensional distributions. With each iteration the Gibbs sampler updates a component of the vector θ with a sample from its marginal distribution, conditional on the current value of all other components (Neal 1993, p. 47). Let $\pi(\theta, D), \theta \in S \subseteq \Re^p$ denote the joint posterior distribution with θ_p being the p's component of θ that has a conditional posterior distribution:

$$\pi_p(\theta_p|\theta_1, \theta_2, ..., \theta_{p-1}, \theta_{p+1}, ..., \theta_p, D) \tag{4.7}$$

Gibbs sampling uses the following steps to approximate the joint posterior by sequentially updating the components of θ (Chib and Greenberg 1996):

1. Set $i = 0$ and initialize θ^0 with a set of values: $\theta^0 = (\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_p^{(0)})$
2. Sample

$$\begin{array}{lll}
\theta_1^{(i+1)} & \text{from} & \pi_1(\theta_1|\theta_2^{(i)}, \theta_3^{(i)}, ..., \theta_p^{(i)}, D) \\
\theta_2^{(i+1)} & \text{from} & \pi_2(\theta_2|\theta_1^{(i+1)}, \theta_3^{(i)}, ..., \theta_p^{(i)}, D) \\
\theta_3^{(i+1)} & \text{from} & \pi_3(\theta_3|\theta_1^{(i+1)}, \theta_2^{(i+1)}, \theta_4^{(i)} ..., \theta_p^{(i)}, D) \\
... & & \\
\theta_p^{(i+1)} & \text{from} & \pi_p(\theta_p|\theta_1^{(i+1)}, \theta_2^{(i+1)}, ..., \theta_{p-1}^{(i+1)}, D)
\end{array} \tag{4.8}$$

3. Set $i = i + 1$ and repeat from step 2 until the a predetermined number of iterations is reached.

The number of iterations it takes the Markov Chain to overcome the influence of the initial starting values $\theta^{(0)}$ before converging to the target distribution is called the "burn-in" period. If the chain converged, the samples reflect the posterior distribution and can be used to summarize the target density by graphical means or summary statistics. For example, the expectation of a parameter of interest can easily be approximated by taking the average over the samples drawn for that parameter. Under general conditions, the ergodicity of the Markov Chain guarantees that this estimate is consistent and satisfies the central limit theorem as the number of samples goes to infinity (Chib and Greenberg 1996).

Another method often used to sample from non-conditionally conjugate distributions is the Metropolis Hastings (MH) algorithm (Metropolis et al. 1953; Hastings 1970). Gibbs sampling can be viewed as a special case of this more general procedure (Gelman et al. 2003, p. 293). The MH algorithm generates the next candidate value of the Markov Chain from a proposal density, which is then accepted with probability α or rejected with probability $(1-\alpha)$. The acceptance probability is the relative ratio between the target density at the candidate value to the target density at the current value (Ntzoufras 2009, p. 43):

1. Set $i = 1$ and initialize θ^0 with a set of values.
2. Set $\theta = \theta^{(i-1)}$
3. Sample candidate parameter value θ' from a proposal distribution[9] $Q(\theta'|\theta)$
4. Calculate acceptance probability α:

$$\begin{aligned} \alpha &= \min(1, \frac{P(\theta'|D)Q(\theta|\theta')}{P(\theta|D)Q(\theta'|\theta)}) && (4.9) \\ &= \min(1, \frac{P(D|\theta')\pi(\theta')Q(\theta|\theta')}{P(D|\theta)\pi(\theta)Q(\theta'|\theta)}) && (4.10) \end{aligned}$$

5. Update $\theta^{(i)} = \theta'$ with probability α; otherwise reject and set $\theta^{(i)} = \theta$

[9] If the proposal distribution is symmetrical, $Q(\theta'|\theta) = Q(\theta|\theta')$, the Metropolis Hastings algorithm collapses into a pure Metropolis algorithm. In this case the acceptance rate is calculated by: $\alpha = \min(1, \frac{P(\theta'|D)}{P(\theta|D)})$.

6. Set $i = i + 1$ and repeat from step 2 until the a predetermined number of iterations is reached.

The MH algorithm will converge to its equilibrium distribution regardless the choice of the proposal distribution Q, if it is positive over the whole support of the target distribution. In practice the proposal acceptance rate α benefits from a carefully chosen and tuned, yet easy to sample, proposal distribution (Gelman et al. 2003, p. 307).

The above description of the algorithm block-updates the complete vector θ, a component wise update, that applies MH to each or several components of θ, is often used within a Gibbs framework. This method is called "Metropolis within Gibbs" and is used to further reduce the dimensionality of the sampling distributions involved (Ntzoufras 2009, p. 45).

Metropolis-Hastings within Gibb already enables the approximation of many target posterior distributions of complex models within the Markov-Chain Monte Carlo framework. Additionally there are a number of sampling techniques that have been developed to improve convergence rates and/or efficiently handle specific distributions. For example, the "slice-sampler" (Neal 1997) is a general procedure to sample from one-dimensional conditional distributions, the "reversible jump sampler" allows trans-dimensional Markov chain simulation (Green 1995), or Hybrid/Hamiltonian Monte Carlo (Brooks et al. 2011, pp. 113-160) replaces the random walk with a momentum directed exploration of the target distribution. Hamilton Monte Carlo sampling can be more efficient and robust for models with complex posteriors (Carpenter et al. 2015). While these methods rest on the foundation that theoretically "long run" convergence is guaranteed, practical considerations still play an important role in establishing if a finite Markov-Chain Monte Carlo simulation sufficiently converged to the target posterior distribution.

4.3 Diagnostics of Markov Chain Convergence

In contrast to maximum likelihood estimation or ordinary least squares methods, which yield point estimates, "convergence" in a Bayesian framework refers to the approximation of target posterior distributions (Congdon 2003, p.18). Most standard econometric models have the desirable properties to ensure theoretic ergodicity of the Markov Chain, so that the simulated Markov Chain converges (Chib and Greenberg 1996). But in practice, the speed of convergence and the assessment if convergence is sufficiently reached are important issues. Generally the rate of convergence is subject to several influences and generally cannot be predetermined (Gelman et al. 2009). For example, highly autocorrelated Markov chains require many iterations to fully navigate the target distribution and the length of the burn-in phase depends on the choice of initial values, the data and the model itself (Braun and Damien 2015). Finite machine precision and practical computational limitations in conjunction with multi-modal posterior distributions may lead to situations where the Gibbs sampler fails to visit all modes, even though theoretically possible (Rossi, Allenby, and McCulloch 2005, p. 100).

Figure 4.1 illustrates different scenarios that can arise in one and two chain MCMC simulations. Generally, a trace plot shows draws from the target marginal posterior distribution for a parameter (y-axis) over a number of iterations (x-axis). The first two trace plots (Figure 4.1 a and b) depict single chain MCMC simulations for the same parameter. The shape of trace plot (a) indicates that the range of values drawn in the first iterations is different from that of the last iterations. Also with later iterations the average value drawn seems to visibly increase, forming a trend. These aspects are a clear sign that the chain did not converge in the first 2,000 iterations. The trace plot in Figure 4.1 (b) shows the same MCMC simulation from iteration 3,000 to 8,000. Here the trace plot resembles a "caterpillar" pattern, that exhibits no visible trends and a stationary range of values and variance. Such a pattern usually indicates convergence.

In some cases though a trace-plot might seem stationary with constant variance over a large number of iterations before the Gibbs sampler starts to explore another region of the posterior. Therefore, it is common practice to use more than one chain to simulate from the Markov chain. Multiple chains with dispersed starting values do not only allow to assess individual chain convergence, but also if the chains converge relative to each other. Figure 4.1 (c) shows a case where two chains are simulated in parallel. The red chain seems relatively stable without

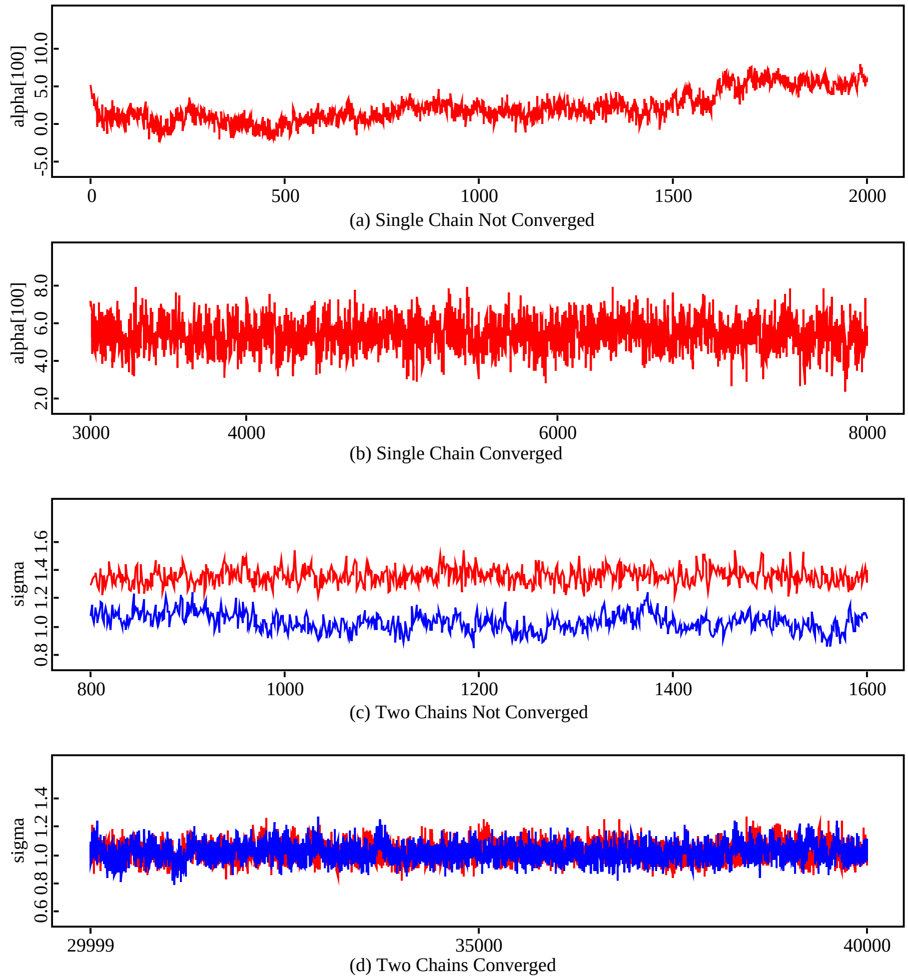

Figure 4.1: Trace Plots of Converged (b+d) and Non-Converged (a+c) Markov Chains for One Chain (a+b) and Two Chain Monte Carlo Simulations (c+d)

trends, and judging from that chain alone one might be mislead to assume convergence. The blue chain however, seems much less stationary and thus, has not yet converged. Much more importantly, though, the two chains have not converged relative to each other. If inferences would be made from one of the chains it would yield different results than those made from the other — clearly this should not be the case for Markov chains from the same posterior target distribution. Finally, Figure 4.1 (d) shows a much later stage of the simulation over a larger number of iterations. Here, the influence of the dispersed starting values have been overcome and both chains have converged individually and relative to each other. It is recommended practice to inspect the posterior trace plots to assess convergence (Ntzoufras 2009; Jen, Chou, and Allenby 2009; Abe 2009; Rossi, Allenby, and McCulloch 2005; Congdon 2003).

Additionally, a number of numerical metrics have been developed to assess the degree to which a MCMC simulations has converged towards the target distribution (e.g., Cowles and Carlin 1996; Brooks and Gelman 1998). The Gelman and Rubin method is based on multiple-chain sampling from dispersed starting values and relates the within-chain variation to the between-chain variation (Gelman and Rubin 1992). It operates on the marginal distributions of the scalar parameters of interest and uses an analysis of variance (ANOVA) approach to estimate the pooled posterior variance $\hat{V}$ from the between- and within-chain variability (Ntzoufras 2009, p. 144):

$$\hat{V} \quad = \quad \frac{T'-1}{T'} WSS + \frac{BSS}{T'} \frac{\kappa+1}{\kappa} \tag{4.11}$$

Here, κ denotes the number of chains, T' the number of iterations in each chain, BSS/T' is the between-chain variance and WSS is the mean of the within-chain variances.

$$\hat{R} \quad = \quad \frac{\hat{V}}{WSS} = \frac{T'-1}{T'} + \frac{BSS/T'}{WSS} \frac{\kappa+1}{\kappa} \tag{4.12}$$

Because, the real variance of the posterior is unknown WSS is used as an (under) estimate, so that $\hat{V}/WSS$, is an (over) estimate of the scale[10] reduction factor $\hat{R}$ (Gelman and Rubin 1992). As the MCMC simulation converges to its target distribution, and if the number of iterations is large enough, $\hat{R}$ converges to 1. This idea was refined by Brooks and Gelman (1998) to account for the sampling variability in the variance estimates. If the estimated degrees of freedom for the

[10] Technically it is an variance reduction factor.

pooled posterior variance $\hat{V}$ are denoted by d, then:

$$\hat{R}_c = \frac{d+3}{d+1}\hat{R} \tag{4.13}$$

This Brooks-Gelman-Rubin (BGR) statistic indicates convergence if the BGR ratio $\hat{R}_c < 1.1$ (Brooks and Gelman 1998; Carpenter et al. 2015). It is the most widely used statistic to assess convergence. The concept of "effective sample size" is derived from the BGR approach and estimates the number of independent samples (Gelman et al. 2003, p. 298):

$$n_{eff} = \kappa T'(\hat{V}/BSS) \tag{4.14}$$

For example, highly autocorrelated chains tend to have small effective sample size relative to the number of iterations. It can also be useful to have an estimate of the burn-in time before sampling (Chib and Greenberg 1996), for example, by analyzing the rate of convergence of the Markov Chain with the target density Raftery and Lewis (1992).

4.4 Measures of Forecast Accuracy

4.4.1 Measures of Forecast Accuracy and Aggregation

The measurement of predictive performance is based on a number of error metrics commonly used in literature. Generally, in this thesis, the observations are split into two subsamples: a training sample and a hold-out sample. The hold-out data is only used for validation of model forecasts, while the learning sample is used to estimate the model parameters. Let x_j denote the hold-out observation at time j, let f_j denote the model forecast for time j, and let F denote the length of the hold-out period. With the definition of the errors as $e_j = x_j - f_j$, the error metrics are as follows (Hyndman and Koehler 2006):

$$MSE = \text{Mean Square Error} = \sum_{j=1}^{F} e_j^2/F \tag{4.15}$$

$$RMSE = \text{Root Mean Square Error} = \sqrt{MSE} \tag{4.16}$$

$$MAE = \text{Mean Absolute Error} = \sum_{j=1}^{F} |e_j|/F \tag{4.17}$$

$$MAPE = \text{Mean Absolute Percentage Error} = \frac{100}{F}\sum_{j=1}^{F}\left|\frac{e_j}{x_j}\right| \tag{4.18}$$

In forecast literature, the *MSE* is also referred to as "average squared predictor error" (ASPE), *RMSE* is sometimes referred to as "rooted average squared predictor error" (RASPE), and *MAE* is also called "mean absolute deviation" (MAD) (Leeflang et al. 2000).

The main focus of this thesis are individual level forecasts of purchase frequencies. In order to obtain a measure of accuracy for the full customer base of a retailer the individual level errors have to be aggregated into an overall measure. Thus, I compute the *MSE* and *RMSE* for every customer and aggregate by mean and median. *MAE* and *MAPE* are calculated irrespective of the customer they belong to, over the whole customer base. *MAE* is used for forecasts of purchase frequencies as *MAPE* is not defined for observed values of zero. In case of the forecast of interpurchase times, *MAPE* is used.

4.4.2 Classification and Receiver Operating Characteristic

The dichotomous prediction of customer inactivity (or activity) is essentially a classification task. In managerial decision making, it is often not enough to just assign a probability or measure that indicates the likelihood of a customer becoming inactive (or equivalently staying active). Instead, a manager faces a clear-cut decision if, for example, to offer specific customers incentives, or not. For this reason a cutoff at a fixed threshold determined by the managerial task at hand has to be made. This is not only true for probabilistic models, but also for heuristic approaches. For example the hiatus heuristic, which I use as a baseline, depends on choosing a cutoff value for the length of the hiatus. If the length of the hiatus, the time that elapsed since the last purchase, is longer than that predetermined threshold a customer is classified inactive.

A test that would only assess a models performance at typical or so-called "natural" cutoff values — for example, a 50% or 95% certainty level for a probability model or a 4-month cutoff for the hiatus heuristic — would be misleading, because the true performance of a classifier can only be assessed over the whole range of cutoff values. The receiver operating characteristic (ROC) addresses this issue, as it plots the performance of a classifier over the whole range of its potential cutoff values.

A ROC analysis is a two dimensional graph in which the true positive rate is plotted on the Y axis and the false positive rate is plotted on the X axis for all possible thresholds (Fawcett 2006). If for example the task is to classify customers ac-

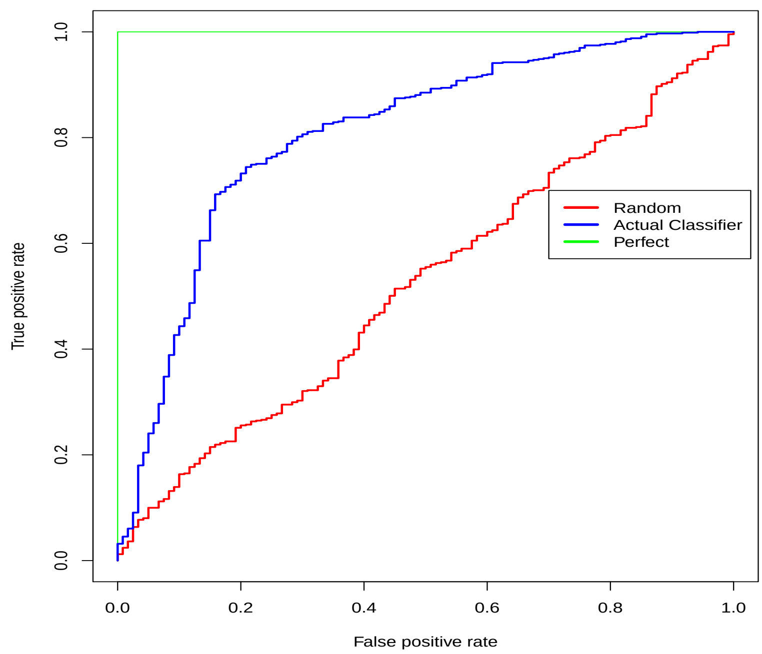

Figure 4.2: Receiver Operating Characteristic of a Random Classifier, an Actual Classifier and a Perfect Classifier

cording to their likelihood of remaining inactive in the future (the positive) then a true positive is a customer who is classified as inactive and who remained inactive during the testing period.

The number of true positives divided by the total number of positives is the true positive rate, while the number of false positives divided by total number of negatives, is the false positive rate (Fawcett 2006). A perfect classifier would have a true positive rate of one and a false positive rate of 0. This implies that the perfect ROC curves goes from $(0,0)$ to $(0,1)$, and then to $(1,1)$ — compare the green line of a perfect classifier in Figure 4.2. ROC curves along the Diagonal $(0,0)-(1,1)$ represent random guessing — see the red line in Figure 4.2. Any classifier worse than random guessing, which implies going through the lower right triangle, can be improved upon by inverting the measure, mirroring it to the left-upper triangle. The better the classifier the further out in the left upper triangle the ROC curve will be — for example in Figure 4.2 the green perfect classifier is further out in the upper left triangle than the blue actual classifier.

Chapter 5

Research Data

5.1 Empirical Settings and Datasets

This thesis focuses on forecasting individual customer purchase behavior in noncontractual retail settings. I used three large-scale datasets from distinct and diverse retail settings to assess the forecast accuracy of the proposed models. The data stem from a German DIY retail chain, a German apparel retail chain and an online DVD/CD mail-order retailer (CDNOW). In all three cases the buyer-seller relationship is noncontractual.

The DIY retailer provided 76,260 transactional data points from 2,460 customers over a 31-month period beginning in August 2003 and ending in February 2006. The product range includes, for example, power tools, paint, hardware and accessories. The dataset from the apparel retailer contains transactions from customers spanning a time period of two years. The data recorded 337,321 items purchased from January 2002 until the end of December 2003 by 27,814 customers. The product categories cover different types of apparel for both men and women, ranging from low-price convenience clothing to high-price business attire.

As an additional reference, I include a publicly available dataset provided by Bruce Hardie (2011). This data set describes the purchase histories of 23,570 customers of the online retailer CDNOW from January 1997 to June 1998 and has been used multiple times in previous research (e.g., Fader, Hardie, and Lee 2005b; Glady, Lemmens, and Croux 2015; Jerath, Fader, and Hardie 2011; Wübben and von Wangenheim 2008). I use the $1/10^{\text{th}}$ subsample containing data on a cohort of 2,357 customers that has been made available for download from Bruce Hardie's website (Hardie 2011). The retailer engages in the sale of music CDs and movie DVDs via mail-order exclusively from its website.

5.2 Data Processing for the Dynamic Model of Purchase Timing

I converted the transactional data for each customer into individual level interpurchase times as this is the unit of measurement used by the Dynamic Model of Purchase Timing. An interpurchase time denotes the time elapsed between two consecutive purchases. This operation transforms a customer's n purchase times into $n-1$ interpurchase times with no loss of information if a customer bought at least on two occasions.

In the case of the DVD/CD and the apparel retailer, I aggregated the data in terms of calendar weeks. I refrained from using daily interpurchase times, to mitigate the effects of purchase decisions being split into two different purchase events as recommended by Allenby, Leone, and Jen (1999). This issue might arise if a customer returns a good purchased a day earlier and during the same visit buys a substitute product. In this case, one decision to purchase an item would register as two purchase events, resulting in an additional interpurchase time. By choosing weekly aggregation, this effect is reduced. For the DIY retailer, I calculated monthly interpurchase times, because the data did not allow for a finer granularity.

In order to ensure a true cohort of customers I included only those customers in the analysis that made their first purchase within the first month of data. Due to the DMPT's use of lagged covariates in the temporal dynamic part of the model (Venkatesan, Kumar, and Bohling 2007), only customers with at least

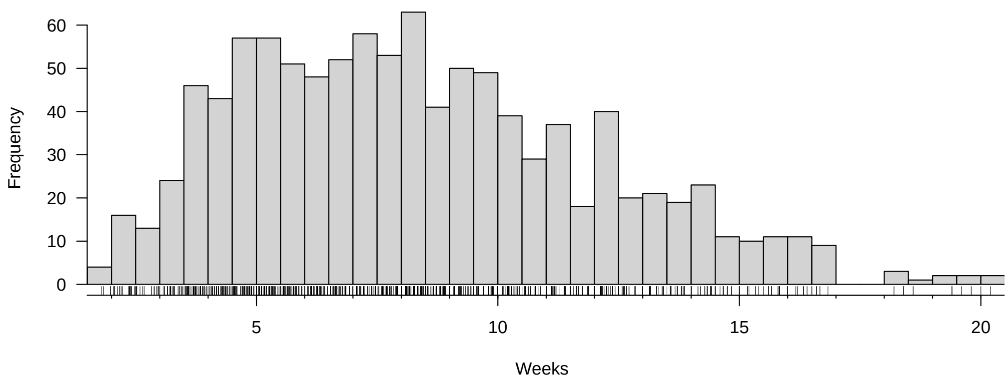

Figure 5.1: Frequency of Average Interpurchase Times — Apparel

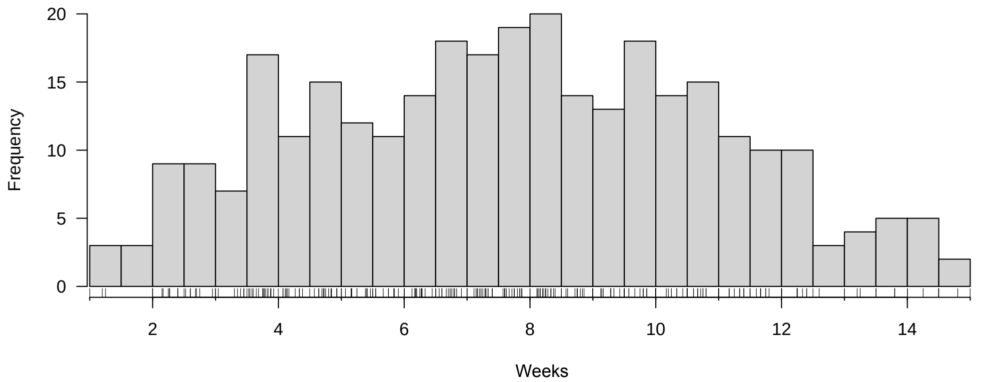

Figure 5.2: Frequency of Average Interpurchase Times — CDNOW

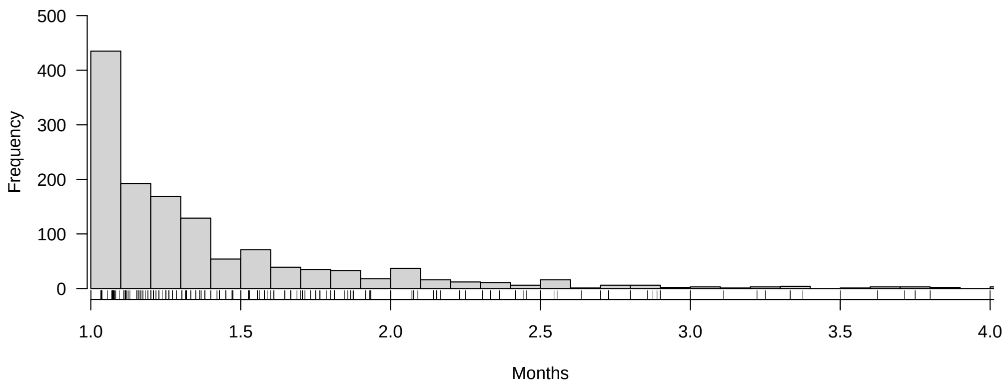

Figure 5.3: Frequency of Average Interpurchase Times — DIY

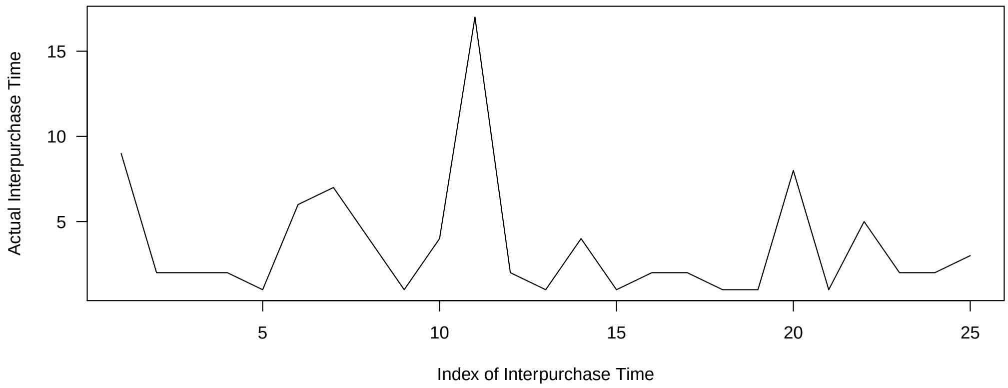

Figure 5.4: Interpurchase History of a Single Customer — Apparel

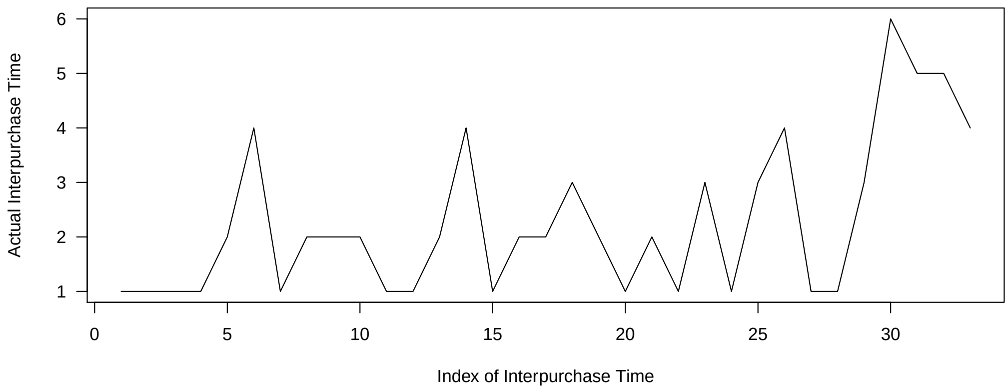

Figure 5.5: Interpurchase History of a Single Customer — CDNOW

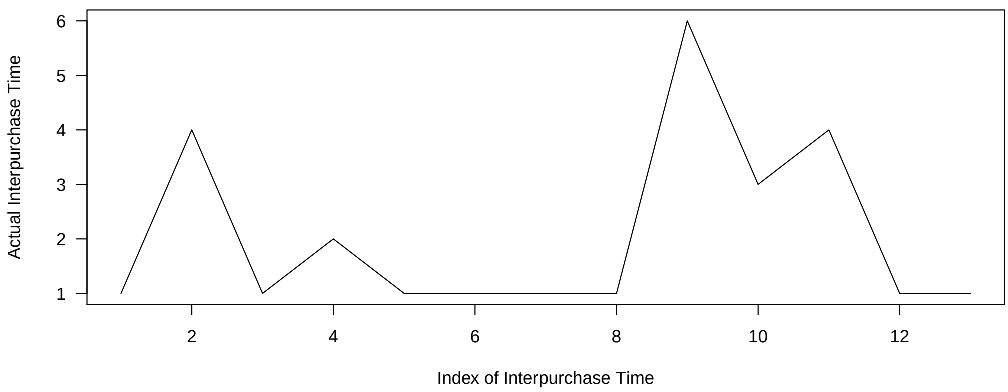

Figure 5.6: Interpurchase History of a Single Customer — DIY

four recorded interpurchase times in the learning period are selected for the final set.

For the apparel retailer, this leaves 1,033 customers with 100 weeks of data that are split into two parts: the first 60 weeks for the learning sample and 40 weeks for the hold-out sample. In the case of the DIY retailer, this results in 31 months of data that include a 21-month learning period and ten months of hold-out data for 1,316 customers. The DVD/CD retailer data yielded 309 customers with 78 weeks of data, of which 62 weeks are used for learning and 16 are reserved as hold-out.

The frequency histograms of the average interpurchase times of all customers for each of the three datasets are depicted in Figures 5.1, 5.2 and 5.3. Figures 5.4, 5.5 and 5.6 contain the data of one individual customer (randomly chosen) for each of the three datasets. These graphs illustrate a customer's complete interpurchase history. The spikes are periods of inactivity that the retailers want to avoid — for example, by providing incentives to customers who might potentially become inactive.

5.3 Data Processing for the Hierarchical Seasonal Effects Models

I specified the Hierarchical Bayesian Seasonal Effects Models in terms of purchases per time unit instead of interpurchase times. Also, these models, unlike the DMPT, are not restricted to customers with a minimum of four purchases.

In order to ensure a true cohort of customers I included only those customers in the analysis that made their first purchase within the first month of available data. The granularity of the DIY retailer's customers' monthly purchase data was taken as-is. In the case of the DVD/CD and the apparel retailers, I aggregated the customer data in terms of months from their individual time-stamped purchases. Monthly data (twelve seasonal components) reduce the computational demands of weekly (52 seasonal components) or daily (365 daily components) granularity

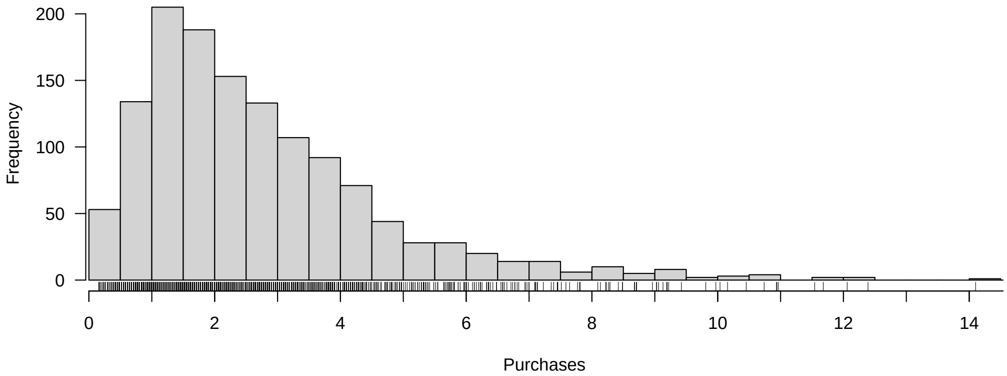

Figure 5.7: Frequency of Average Number of Purchases — DIY

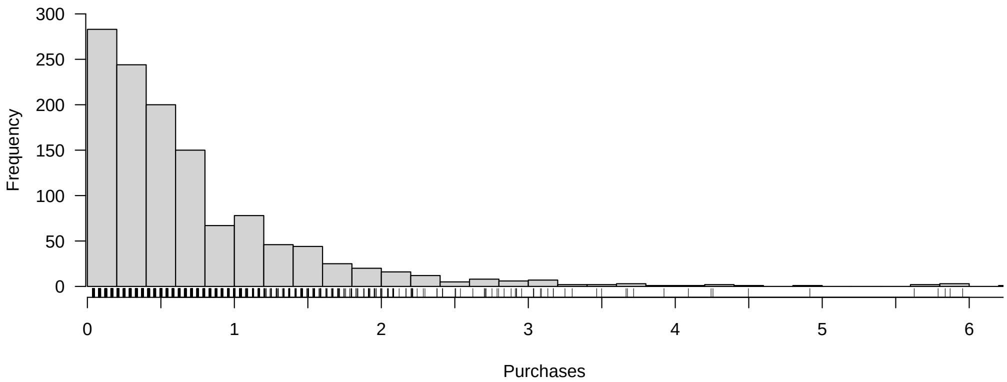

Figure 5.8: Frequency of Average Number of Purchases — Apparel

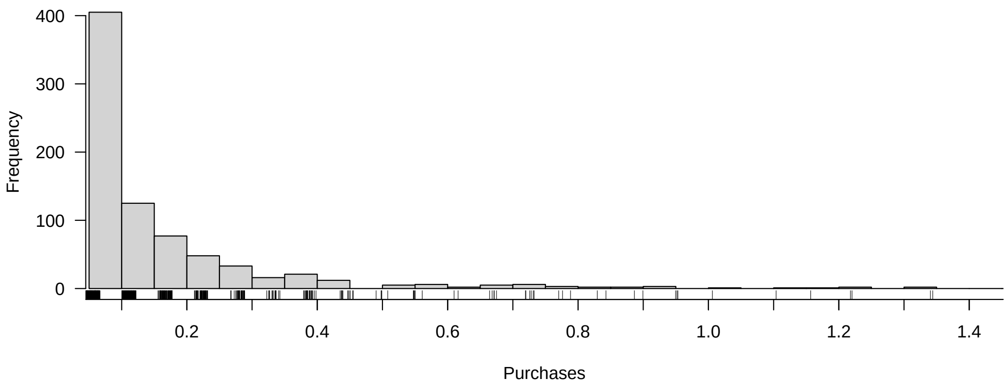

Figure 5.9: Frequency of Average Number of Purchases — CDNOW

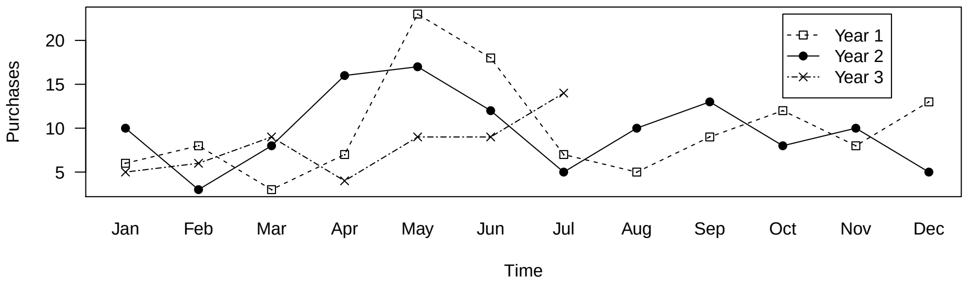

Figure 5.10: Purchases of a Single Customer — DIY

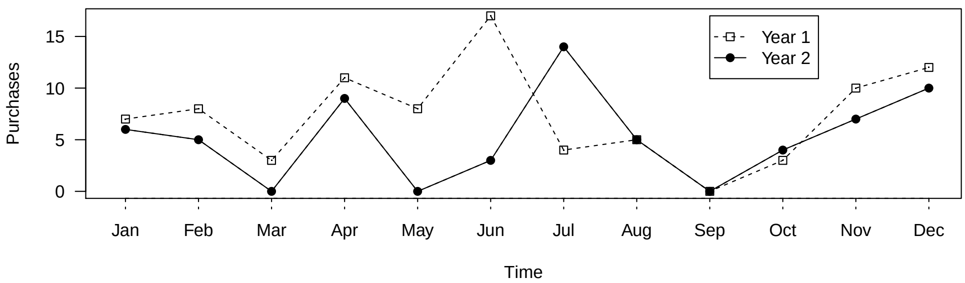

Figure 5.11: Purchases of a Single Customer — Apparel

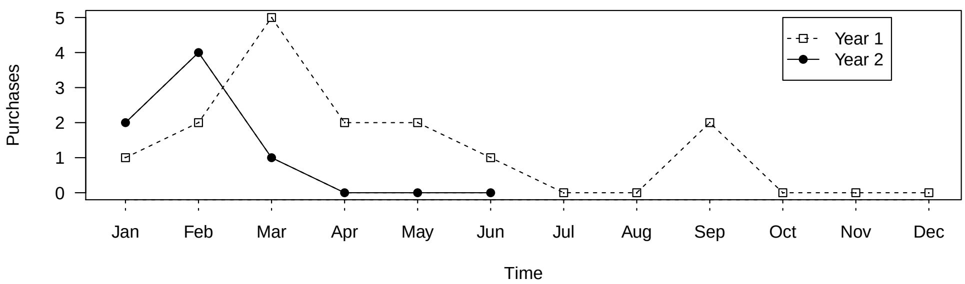

Figure 5.12: Purchases of a Single Customer — CDNOW

substantially. I reserved four month worth of data as hold-out.

In the case of the DIY retailer, this procedure results in 31 months of data that include a 27-month learning period and four months of hold-out data for 1,329 customers. For the apparel retailer, this leaves 1,230 customers with 24 months of data that are split into two parts: the first 20 months for the learning sample and the last four months for the hold-out sample. The DVD/CD retailer data yielded 781 customers with 18 months of data, of which 14 months are used for learning and four months are reserved as hold-out. The frequency histograms of the customers' average number of purchases for each of the three datasets are depicted in Figures 5.7, 5.8, and 5.9.

Figures 5.10, 5.11, and 5.12 illustrate the resulting data for one randomly chosen individual customer from each dataset. The graphs depict the number of monthly purchases the customer made over time with successive years being shown in overlay. This illustrates the seasonal nature of the data and the relative small amount of seasonal overlap in the datasets. For example, the CDNOW dataset features a total of only 18 months of data. With a four month hold-out period this leaves 14 months of data with only a two months overlap between each year as a sample for the model to learn a seasonal structure.

Chapter 6

Dynamic Model of Purchase Timing and Customer Inactivity

This chapter comprises the development of an improved rule for the classic "Dynamic Model of Purchase Timing" (DMPT) by Allenby, Leone, and Jen (1999), a receiver operating characteristic analysis of the classifications results, and a benchmark against managerial heuristics. First, I will recap the original model by Allenby and colleagues (1999), then I will discuss methods of deriving measures of customer inactivity and propose a refined Bayesian estimate of customers' behavioral consistency with the model. Finally, I assess and discuss the predictive performance of both the original and refined model in relation to baseline managerial heuristics.

6.1 General Model Framework

Allenby, Leone, and Jen (1999) proposed the DMPT to address the cross-sectional heterogeneity of customers and allow individual disaggregate forecasts of customer behavior. The DMPT augments a hierarchical structure that relates individual interpurchase times and cross-sectional heterogeneity by a component mixture with temporal dynamics. For clarity, I present the model in four steps that build upon each other and together form the complete model:

1. the hierarchical random-effects generalized gamma model;
2. the generalized gamma component mixture;
3. temporal dynamics and link function; and
4. prior distributions and hyperparameters.

6.1.1 Hierarchical Random-Effects Generalized Gamma Model

The DMPT, in contrast to the HSM and HSMDO, operates on interpurchase times instead of purchase frequencies per time period. In this and the following subsections (6.1.1-6.1.4) I follow the original model description by Allenby, Leone, and Jen (1999) with a slightly adapted notation. In the one component case, if t_{ij} is the j^{th} interpurchase time for customer i, then t_{ij} is distributed generalized gamma:

$$t_{ij} \sim GG(\alpha, \lambda_i, \gamma) \overset{\text{p.d.f.}}{\rightleftharpoons} \frac{\gamma}{\Gamma(\alpha)\lambda_i^{\alpha\gamma}} t_{ij}^{\alpha\gamma-1} e^{-(t_{ij}/\lambda_i)^{\gamma}} \tag{6.1}$$

The variables α and γ are overall group level parameters, while λ_i is an customer specific individual parameter. The expected value of the generalized gamma distribution using this parameterization is given by:

$$E[GG(\alpha, \lambda_i, \gamma)] = \frac{\Gamma(\alpha + 1/\gamma)}{\Gamma(\alpha)} \lambda_i \tag{6.2}$$

Assuming that there is only one component, the heterogeneity across customers is captured by an inverse generalized gamma distribution (IGG) of the λ_i's:

$$\lambda_i \sim IGG(\nu, \theta, \gamma) \overset{\text{p.d.f.}}{\rightleftharpoons} \frac{\gamma}{\Gamma(\nu)\theta^{\nu\gamma}} \lambda_i^{-\nu\gamma-1} e^{-(1/\theta\lambda_i)^{\gamma}} \tag{6.3}$$

The heterogeneity across customers is governed by the overall level parameters ν, θ, and γ of the IGG distribution. The expected value using this parameterization is given by:

$$E[IGG(\nu, \theta, \gamma)] = \frac{\Gamma(\nu - 1/\gamma)}{\Gamma(\nu)\theta} \tag{6.4}$$

6.1.2 Generalized Gamma Component Mixture

The following component mixture extends the one component case to the k component model. If the j^{th} interpurchase time for customer i is a mixture of k generalized gamma components, Equation (6.1) becomes:

$$t_{ij} \sim \sum_k \phi_{ijk} GG(\alpha_k, \lambda_{ik}, \gamma_k) \overset{\text{p.d.f.}}{\rightleftharpoons} \sum_k \phi_{ijk} \frac{\gamma_k}{\Gamma(\alpha_k)\lambda_{ik}^{\alpha_k\gamma_k}} t_{ij}^{\alpha_k\gamma_k - 1} e^{-(t_{ij}/\lambda_{ik})^{\gamma_k}} \tag{6.5}$$

Each of the k components has it's set of parameters α_k, λ_{ik} and, γ_k. The individual level parameters λ_{ik} follow k IGG distributions with parameters ν_k, θ_k, and γ_k:

$$\lambda_{ik} \sim IGG(\nu_k, \theta_k, \gamma_k) \overset{\text{p.d.f.}}{\rightleftharpoons} \frac{\gamma_k}{\Gamma(\nu_k)\theta_k^{\nu_k\gamma_k}} \lambda_{ik}^{-\nu_k\gamma_k - 1} e^{-(1/\theta_k\lambda_{ik})^{\gamma_k}} \tag{6.6}$$

The link function ϕ determines the mixture proportions of the k generalized gamma distributions. To allow more abrupt changes in the predicted interpurchase times, ϕ is dependent on j through time-varying covariates that will be explained in the next step.

6.1.3 Temporal Dynamics and Link Function

For $k = 3$, Allenby, Leone, and Jen (1999) specified ϕ so that the three components correspond to three levels of purchase frequency. The first component is stationary for every individual i over all interpurchase times j and reflects the probability of an individual being in a "super-active" state of purchasing:

$$\phi_{ij1} = 1 - \Phi(\beta_{01i}) \quad \text{(super-active state)} \tag{6.7}$$

Φ denotes the cumulative normal distribution. The resulting link function ϕ_{ij1} is the probability that customer i at the j^{th} interpurchase time is in state 1, similar to a traditional multinomial probit model. The other components model active and inactive[11] purchase behavior and are tied to time-varying covariates as fol-

[11] "Inactive" is the original term used by Allenby, Leone, and Jen (1999). A more precise description would be "least active" state.

lows:

$$\phi_{ij2} = \Phi(\beta_{01i})(1-\Phi(x'_{ij}\beta_{2i}+\beta_{02i})) \qquad \text{(active state)} \tag{6.8}$$

$$\phi_{ij3} = \Phi(\beta_{01i})\Phi(x'_{ij}\beta_{2i}+\beta_{02i}) \qquad \text{(inactive state)} \tag{6.9}$$

$$x'_{ij} = (\log(t_{ij-1}), \log(t_{ij-2}), \log(t_{ij-3}))' \tag{6.10}$$

The vector x'_{ij} consists of three lagged interpurchase times that serve as time-varying covariates. The vector β_{2i} has three elements that reflect the coefficients for the three lagged interpurchase times. The intercept term of this submodel is β_{02i}. The strength of the relationship between the lagged interpurchase times and the state probabilities is allowed to vary between customers. Therefore, a random-effects specification completes the model:

$$\beta'_i = (\beta_{01i}, \beta_{02i}, \beta'_{2i}) \tag{6.11}$$

$$\beta_i \sim MVN(\bar{\beta}, V) \overset{\text{p.d.f.}}{\rightleftharpoons} (2\pi)^{-d/2}|V|^{-1/2}\exp[-\tfrac{1}{2}(\beta_i-\bar{\beta})'V^{-1}(\beta_i-\bar{\beta})] \tag{6.12}$$

MVN is the multivariate normal distribution with the variance-covariance matrix V, a positive definite matrix of size $d \times d$ with $d = 5$.

6.1.4 Prior Distributions and Hyperparameters

The Bayesian approach requires the specification of prior distributions for parameters: $\alpha_k, \nu_k, \theta_k, \bar{\beta}, V$, and γ_k. The subscript k denotes the different mixture components. For identification of the mixture model order is imposed on θ_k, so that $\theta_1 > \theta_2 > ... > \theta_k$. The parameters g and G have been set to $g = 15$ and $G = 15\mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix. IW denotes the inverted Wishart distribution, which is the natural conjugate prior for the variance-covariance matrix of a multivariate normal distribution. The parameterization of g and G ensures a relatively diffuse prior (Rossi, Allenby, and McCulloch 2005, p. 30). The definition of priors, then, is:

$$\theta_k \sim IGG(a_0, b_0, \gamma_k) \tag{6.13}$$

$$V \sim IW(G, g) \tag{6.14}$$

$$\bar{\beta} \sim MVN(\vec{0}, V) \tag{6.15}$$

The values for a_0 and b_0 are set to 10. With $\nu_k, \alpha_k \sim Uniform$, $\gamma_1 = 1.0, \gamma_2 = 1.4$, and $\gamma_3 = 0.8$, the prior specification is complete.

6.2 Refined Rule for Forecasting Customer Inactivity

In noncontractual settings one cannot observe customer defection directly through expiration or cancellation of an ongoing contract. The DMPT model in its original form yields predictions on expected interpurchase times, but does not directly provide information on customers becoming inactive as it lacks an explicit customer lifetime model.

Typically, a customer relationship manager wants to decide at a point in time, after observing j interpurchase times, whether a customer i is still active or not. Allenby, Leone, and Jen (1999) suggest classifying a customer as inactive if the right-censored spell r_i (recency) — the time that elapsed since the last purchase — is inconsistent with the model's prediction. Accordingly, Reinartz, Thomas, and Kumar (2005) consider a customer still active if the expected time until the next purchase $t_{i,j+1}$ exceeds the time elapsed since the last purchase r_i; otherwise, the customer is considered inactive. However, in this form, the rule does not account for the variance of the prediction.

Another, more conservative cutoff rule proposed by Wu and Chen (2000a) applies only if r_i exceeds the predicted interpurchase time by three standard deviations. If this is the case the customer is deemed to be permanently inactive. This approximation introduces bias, because the predictive posterior distribution is not guaranteed to be normal, it might not even be symmetrical.

I improve upon these rules by using the Bayesian structure of the model to directly compute the probability of the prediction exceeding the right-censored spell r_i. The MCMC simulation yields the full posterior distribution of the predicted interpurchase times and thus, includes all moments such as expectation, variance, and skewness. This allows me to compute the exact probability of model inconsistency with the right censored spell of inactivity. This probability can be used to impose any desired level of confidence on the drop-out of a customer.

The second improvement addresses a fundamental issue that arises because of the finite length of the hold-out period. If the predicted interpurchase time $t_{i,j+1}$ would lead to a purchase well after the hold-out period, the above rules imply that this behavior is model consistent. Clearly, if the predicted interpurchase time indicates a purchase beyond $r+h$, it must also be larger than r. In this case the customer would be classified as active, but one would still expect him to make no purchases during the hold-out period. This introduces bias as making no purchases during the (finite) hold-out period is used as a proxy for customer inactivity.

In order to avoid this bias, I propose a measure that takes into account both causes for observing no purchases during the hold-out period: (1) model inconsistent behavior $t_{i,j+1} < r_i$ and (2) model consistent behavior $t_{i,j+1} > r_i + h$. Given the length of the hold-out period h, the elapsed time since the last purchase r_i, and the predicted interpurchase time $t_{i,j+1}$, the probability that customer i makes no purchases during hold-out is as follows[12]:

$$P_i^{inactive} = P(t_{i,j+1} < r_i \vee t_{i,j+1} > r_i + h) = 1 - P_i^{active}(r_i \leq t_{i,j+1} \leq r_i + h) \quad (6.16)$$

$$P_i^{active}(r_i \leq t_{i,j+1} \leq r_i + h) = \int_{r_i}^{r_i+h} \sum_{k=1}^{K} \phi_{i,j,k} GG(t_{i,j+1}|\alpha_k, \lambda_{ik}, \gamma_k) dt_{i,j+1} \quad (6.17)$$

The calculation of the integral is straightforward in a Bayesian framework, because the MCMC procedure allows one to directly sample a prediction for $t_{i,j+1}$ in each iteration. After the Markov chain converges, one can evaluate the indicator (or step) function with the condition $r_i \leq t_{i,j+1} \leq r_i + h$ for each sample of $t_{i,j+1}$.

12 The conditioning is omitted for brevity.

The average of the results approximates the above integral in Equation (6.17). If $t^{(n)}$ denotes the n^{th} sample for $t_{i,j+1}$ in the Markov chain, then for sufficiently large N:

$$
\begin{aligned}
&P_i^{active}(r_i \leq t_{i,j+1} \leq r_i + h) \\
&\qquad = \int_{r_i}^{r_i+h} \sum_{k=1}^{K} \phi_{i,j,k} GG(t_{i,j+1}|\alpha_k, \lambda_{ik}, \gamma_k) dt_{i,j+1} \\
&\qquad \approx \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{r_i \leq t^{(n)} \leq r_i + h}
\end{aligned}
\tag{6.18}
$$

The above is a more refined way to derive a measure for model inconsistent customer behavior ($1 - P_i^{active}$) given the purchase history, the right censored spell, and the length of the hold-out period. Still, as the model does not contain an explicit customer lifetime model, I expect only a slight improvement compared to the previous rules for the DMPT. In the course of the benchmark analysis, I compare the receiver operating characteristics of the original and this refined classification rule over all possible cutoff values.

6.3 Data Analysis: Parameter Estimation and Prediction

6.3.1 Convergence and Marginal Posterior Estimates

The training of the DMPT model consists of learning about the main parameters' posterior distributions. Training uses only data from the learning period of each retail dataset, reserving the hold-out period for validation. I obtained the marginal posterior distributions for the model's parameters by a MCMC procedure via Gibbs sampling. The DMPT estimation results are based on MCMC runs with 2 chains starting from dispersed initial values.

The MCMC iteratively converges towards the true posterior distribution of the parameter space and after convergence simulates the true posterior distribution. I assessed the approximate convergence of the algorithm by visually inspecting the trace plots of the posterior samples after burn-in and by evaluating the Brooks, Gelman and Rubin (BGR) statistic (Gelman and Rubin 1992; Brooks and Gelman 1998).

For both CDNOW and the apparel data, the trace plots indicate convergence after 150,000 iterations, because there are no visible "trends" and both chains are sufficiently mixed. In the case of the DIY retailer, 150,000 iterations were not enough to establish convergence. The run was extended for a total of 400,000 iterations. The trace plots show convergence after 350,000 iterations with no visible trends and sufficiently mixed chains.

I validated Markov chain convergence by computing the BGR statistic for the last 50,000 nodes. In all cases, the BGR ratios stayed between 0.98 and 1.02, which indicates that the chains have converged to the true posterior distribution (Gelman and Rubin 1992; Brooks and Gelman 1998; Cowles and Carlin 1996). In addition, I simulated the posterior distribution using multiple single-chain runs with over-dispersed starting values that converged to virtually identical parameter values.

I performed thinning by taking every fifth sample after the burn-in period to save memory. Therefore, the posterior distributions are based on $(50,000/5) \times 2 = 20,000$ samples. The resulting marginal posterior densities for the main parameters α_k, θ_k, and ν_k are shown in Figures 6.1-6.3 for each of the retailers. The full marginal posterior densities for the dynamic effect submodel, the vector $\bar{\beta}$, and the precision matrix $prec.V$ are depicted in the Appendix (A.1–A.6). The estimated summary statistics for the model parameters' marginal posterior distributions, including means, standard deviations, medians, and credibility intervals are compiled in Tables 6.1–6.3.

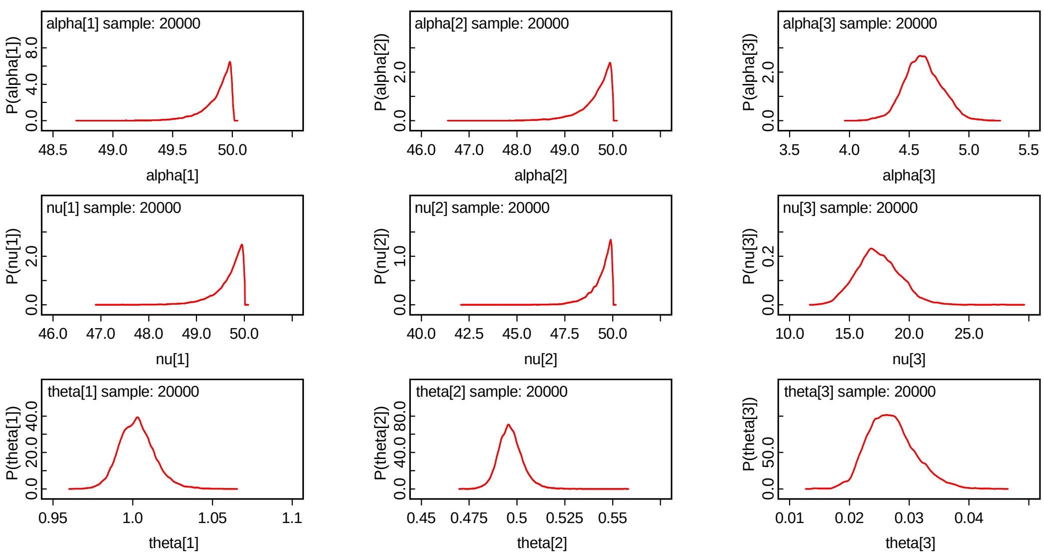

Figure 6.1: DMPT/Apparel Marginal Posterior Densities for α_k, ν_k, and θ_k after 50,000 Iterations per Chain (2 Chains, 150K Burn-In)

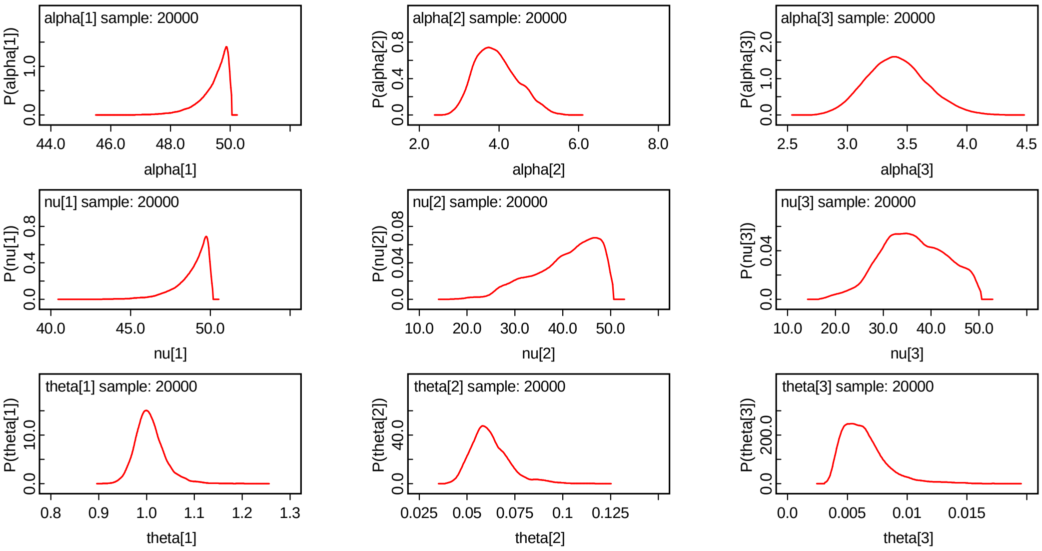

Figure 6.2: DMPT/CDNOW Marginal Posterior Densities for α_k, ν_k, and θ_k after 50,000 Iterations per Chain (2 Chains, 150K Burn-In)

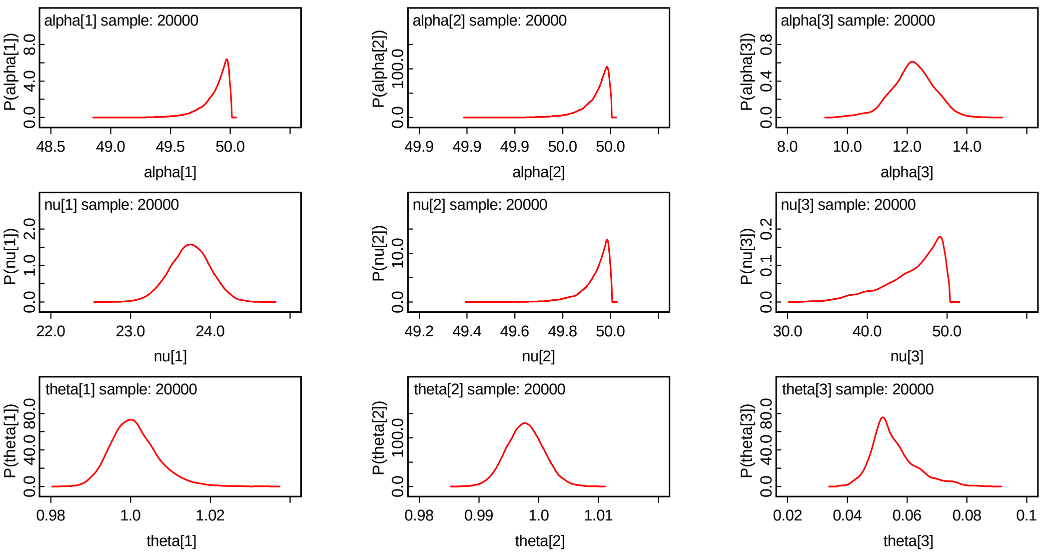

Figure 6.3: DMPT/DIY Marginal Posterior Densities of α_k, ν_k, and θ_k after 50,000 Iterations per Chain (2 Chains, 350K Burn-In)

Table 6.1: DMPT/CDNOW Marginal Posterior Distribution Summaries for MCMC Simulation after 50,000 Iterations per Chain (2 Chains, 150K Burn-In)

Summary of Marginal Posterior Distributions						
node	*mean*	*sd*	*MC error*	*2.5%*	*median*	*97.5%*
alpha[1]	49.460	0.520	0.010	48.070	49.610	49.990
alpha[2]	3.951	0.527	0.033	3.069	3.898	5.071
alpha[3]	3.418	0.250	0.013	2.957	3.409	3.941
nu[1]	48.890	1.078	0.018	46.010	49.220	49.970
nu[2]	40.730	6.673	0.410	26.280	42.120	49.590
nu[3]	36.440	6.762	0.421	23.280	36.150	48.840
theta[1]	1.008	0.031	0.001	0.956	1.004	1.080
theta[2]	0.062	0.010	0.001	0.047	0.061	0.088
theta[3]	0.006	0.002	0.000	0.004	0.006	0.011
bar.beta[1]	1.188	0.090	0.002	1.022	1.185	1.373
bar.beta[2]	−0.358	0.389	0.011	−1.129	−0.358	0.414
bar.beta[3]	0.794	0.344	0.010	0.211	0.762	1.579
bar.beta[4]	0.314	0.264	0.008	−0.179	0.302	0.867
bar.beta[5]	0.730	0.301	0.010	0.188	0.709	1.389
prec.V[1,1]	2.183	0.405	0.008	1.484	2.155	3.060
prec.V[1,2]	−0.245	0.242	0.004	−0.734	−0.240	0.226
prec.V[1,3]	−0.294	0.215	0.003	−0.740	−0.286	0.118
prec.V[1,4]	−0.182	0.235	0.004	−0.655	−0.177	0.277
prec.V[1,5]	−0.168	0.257	0.004	−0.685	−0.165	0.332
prec.V[2,1]	−0.245	0.242	0.004	−0.734	−0.240	0.226
prec.V[2,2]	1.083	0.346	0.010	0.531	1.045	1.873
prec.V[2,3]	0.148	0.240	0.006	−0.334	0.150	0.618
prec.V[2,4]	0.074	0.250	0.006	−0.415	0.071	0.577
prec.V[2,5]	0.164	0.259	0.006	−0.347	0.163	0.680
prec.V[3,1]	−0.294	0.215	0.003	−0.740	−0.286	0.118
prec.V[3,2]	0.148	0.240	0.006	−0.334	0.150	0.618
prec.V[3,3]	0.927	0.329	0.010	0.411	0.878	1.690
prec.V[3,4]	0.009	0.234	0.006	−0.466	0.013	0.462
prec.V[3,5]	0.097	0.251	0.007	−0.410	0.097	0.603
prec.V[4,1]	−0.182	0.235	0.004	−0.655	−0.177	0.277
prec.V[4,2]	0.074	0.250	0.006	−0.415	0.071	0.577
prec.V[4,3]	0.009	0.234	0.006	−0.466	0.013	0.462
prec.V[4,4]	1.096	0.361	0.009	0.520	1.050	1.928
prec.V[4,5]	−0.069	0.251	0.005	−0.579	−0.065	0.420
prec.V[5,1]	−0.168	0.257	0.004	−0.685	−0.165	0.332
prec.V[5,2]	0.164	0.259	0.006	−0.347	0.163	0.680
prec.V[5,3]	0.097	0.251	0.007	−0.410	0.097	0.603
prec.V[5,4]	−0.069	0.251	0.005	−0.579	−0.065	0.420
prec.V[5,5]	1.286	0.388	0.010	0.653	1.249	2.154

Table 6.2: DMPT/Apparel Marginal Posterior Distribution Summaries for MCMC Simulation after 50,000 Iterations per Chain (2 Chains, 150K Burn-In)

Summary of Marginal Posterior Distributions						
node	*mean*	*sd*	*MC error*	*2.5%*	*median*	*97.5%*
alpha[1]	49.870	0.132	0.002	49.520	49.910	50.000
alpha[2]	49.650	0.350	0.007	48.700	49.760	49.990
alpha[3]	4.617	0.157	0.009	4.318	4.610	4.928
nu[1]	49.670	0.327	0.003	48.810	49.770	49.990
nu[2]	49.380	0.612	0.011	47.780	49.570	49.980
nu[3]	17.440	1.844	0.099	14.100	17.330	21.230
theta[1]	1.003	0.011	0.000	0.983	1.002	1.027
theta[2]	0.497	0.006	0.000	0.485	0.496	0.511
theta[3]	0.027	0.004	0.000	0.020	0.027	0.036
bar.beta[1]	1.017	0.034	0.001	0.951	1.017	1.084
bar.beta[2]	1.389	0.190	0.011	1.042	1.378	1.791
bar.beta[3]	0.110	0.086	0.004	−0.062	0.110	0.275
bar.beta[4]	0.199	0.081	0.004	0.044	0.197	0.365
bar.beta[5]	0.254	0.083	0.004	0.091	0.255	0.415
prec.V[1,1]	4.158	0.488	0.017	3.292	4.129	5.194
prec.V[1,2]	−0.626	0.270	0.010	−1.196	−0.612	−0.126
prec.V[1,3]	−0.233	0.305	0.010	−0.841	−0.235	0.379
prec.V[1,4]	−0.532	0.305	0.010	−1.146	−0.528	0.055
prec.V[1,5]	−0.397	0.312	0.011	−1.009	−0.397	0.220
prec.V[2,1]	−0.626	0.270	0.010	−1.196	−0.612	−0.126
prec.V[2,2]	1.734	0.351	0.019	1.125	1.705	2.509
prec.V[2,3]	0.907	0.279	0.012	0.357	0.905	1.465
prec.V[2,4]	0.765	0.275	0.012	0.216	0.760	1.329
prec.V[2,5]	0.726	0.267	0.011	0.201	0.727	1.275
prec.V[3,1]	−0.233	0.305	0.010	−0.841	−0.235	0.379
prec.V[3,2]	0.907	0.279	0.012	0.357	0.905	1.465
prec.V[3,3]	2.823	0.463	0.021	2.003	2.789	3.800
prec.V[3,4]	0.286	0.301	0.012	−0.302	0.287	0.882
prec.V[3,5]	0.750	0.293	0.011	0.198	0.739	1.350
prec.V[4,1]	−0.532	0.305	0.010	−1.146	−0.528	0.055
prec.V[4,2]	0.765	0.275	0.012	0.216	0.760	1.329
prec.V[4,3]	0.286	0.301	0.012	−0.302	0.287	0.882
prec.V[4,4]	2.846	0.482	0.023	1.977	2.817	3.868
prec.V[4,5]	0.815	0.314	0.013	0.228	0.805	1.456
prec.V[5,1]	−0.397	0.312	0.011	−1.009	−0.397	0.220
prec.V[5,2]	0.726	0.267	0.011	0.201	0.727	1.275
prec.V[5,3]	0.750	0.293	0.011	0.198	0.739	1.350
prec.V[5,4]	0.815	0.314	0.013	0.228	0.805	1.456
prec.V[5,5]	2.915	0.481	0.023	2.045	2.882	3.940

Table 6.3: DMPT/DIY Marginal Posterior Distribution Summaries for MCMC Simulation after 50,000 Iterations per Chain (2 Chains, 350K Burn-In)

Summary of Marginal Posterior Distributions						
node	*mean*	*sd*	*MC error*	*2.5%*	*median*	*97.5%*
alpha[1]	49.880	0.116	0.001	49.570	49.920	50.000
alpha[2]	49.990	0.007	0.000	49.970	49.990	50.000
alpha[3]	12.210	0.723	0.048	10.630	12.220	13.560
nu[1]	23.750	0.254	0.005	23.240	23.750	24.230
nu[2]	49.940	0.060	0.000	49.780	49.960	50.000
nu[3]	45.830	3.565	0.218	37.160	46.830	49.880
theta[1]	1.001	0.006	0.000	0.991	1.001	1.015
theta[2]	0.998	0.003	0.000	0.992	0.998	1.004
theta[3]	0.056	0.007	0.000	0.044	0.054	0.074
bar.beta[1]	1.351	0.025	0.001	1.303	1.351	1.400
bar.beta[2]	−2.339	0.092	0.003	−2.533	−2.335	−2.173
bar.beta[3]	0.317	0.150	0.007	0.010	0.321	0.606
bar.beta[4]	0.091	0.158	0.008	−0.230	0.095	0.383
bar.beta[5]	0.503	0.149	0.006	0.191	0.510	0.779
prec.V[1,1]	5.635	0.595	0.013	4.551	5.607	6.871
prec.V[1,2]	1.054	0.317	0.008	0.461	1.047	1.702
prec.V[1,3]	−0.238	0.309	0.007	−0.863	−0.234	0.360
prec.V[1,4]	−0.066	0.308	0.007	−0.682	−0.065	0.540
prec.V[1,5]	−0.056	0.285	0.006	−0.620	−0.055	0.501
prec.V[2,1]	1.054	0.317	0.008	0.461	1.047	1.702
prec.V[2,2]	1.976	0.315	0.011	1.407	1.957	2.650
prec.V[2,3]	0.732	0.267	0.011	0.222	0.722	1.279
prec.V[2,4]	0.625	0.257	0.010	0.151	0.614	1.156
prec.V[2,5]	0.827	0.244	0.009	0.382	0.814	1.341
prec.V[3,1]	−0.238	0.309	0.007	−0.863	−0.234	0.360
prec.V[3,2]	0.732	0.267	0.011	0.222	0.722	1.279
prec.V[3,3]	1.804	0.374	0.014	1.173	1.767	2.638
prec.V[3,4]	0.114	0.236	0.008	−0.364	0.117	0.581
prec.V[3,5]	0.102	0.224	0.007	−0.353	0.107	0.532
prec.V[4,1]	−0.066	0.308	0.007	−0.682	−0.065	0.540
prec.V[4,2]	0.625	0.257	0.010	0.151	0.614	1.156
prec.V[4,3]	0.114	0.236	0.008	−0.364	0.117	0.581
prec.V[4,4]	1.728	0.385	0.015	1.091	1.693	2.574
prec.V[4,5]	−0.090	0.222	0.007	−0.568	−0.078	0.316
prec.V[5,1]	−0.056	0.285	0.006	−0.620	−0.055	0.501
prec.V[5,2]	0.827	0.244	0.009	0.382	0.814	1.341
prec.V[5,3]	0.102	0.224	0.007	−0.353	0.107	0.532
prec.V[5,4]	−0.090	0.222	0.007	−0.568	−0.078	0.316
prec.V[5,5]	1.529	0.319	0.012	0.989	1.500	2.229

6.3.2 Results and Forecast Accuracy

6.3.2.1 Prediction of Customer Inactivity

Predicting whether a customer dropped out and thus remains inactive is a classification[13] task. The receiver operating characteristic (ROC) analysis shows (1) that the refined rule dominates the original rule over all cutoff values and (2) that the results in comparison to the heuristic are mixed with neither method setting itself apart. The refined rule within the DMPT performs better on the DIY and CDNOW data, while the hiatus heuristic performs relatively best on the apparel data. However, the ROC curves suggests that all classification methods perform relative poorly for the apparel retailer's data. This overall picture is corroborated by the area under the curve (AUC) metric and at "natural" cutoff values.

The details of the ROC analysis are shown in Figures 6.4-6.6 for each of the three

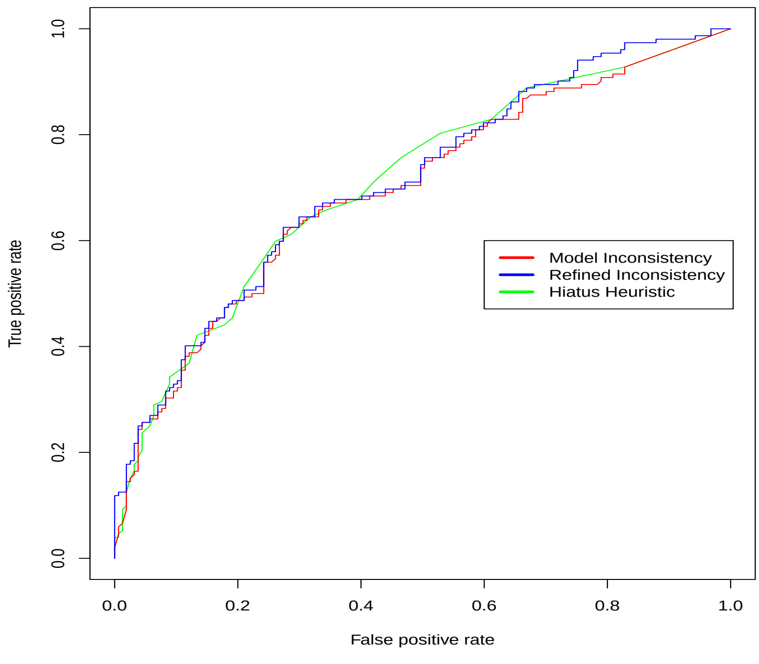

Figure 6.4: DMPT/CDNOW ROC Analysis of Predicting Customer Inactivity with Model Inconsistency and Hiatus Heuristic

[13] For an explanation of the ROC analysis and the classification metrics see Section 4.4.2. The respective hiatus heuristic is described in Section 3.3.2.

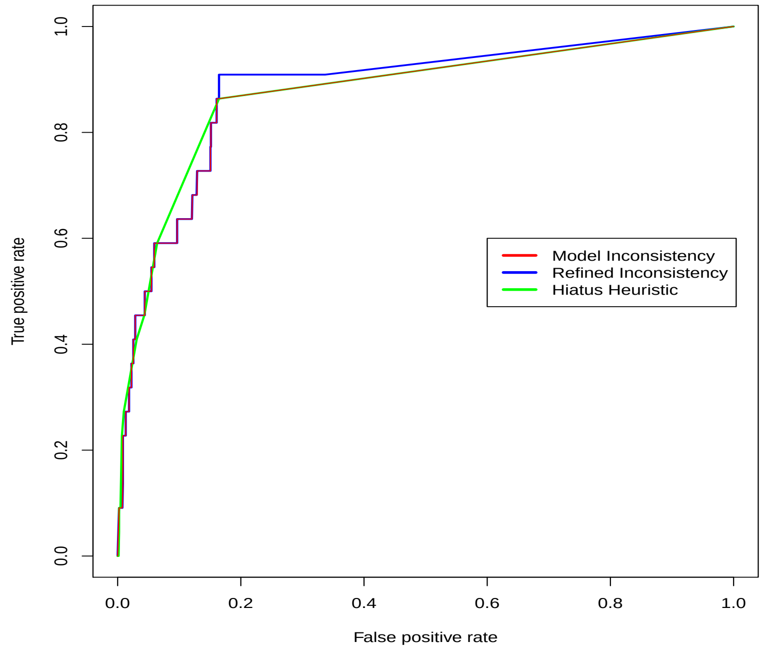

Figure 6.5: DMPT/DIY ROC Analysis of Predicting Customer Inactivity with Model Inconsistency and Hiatus Heuristic

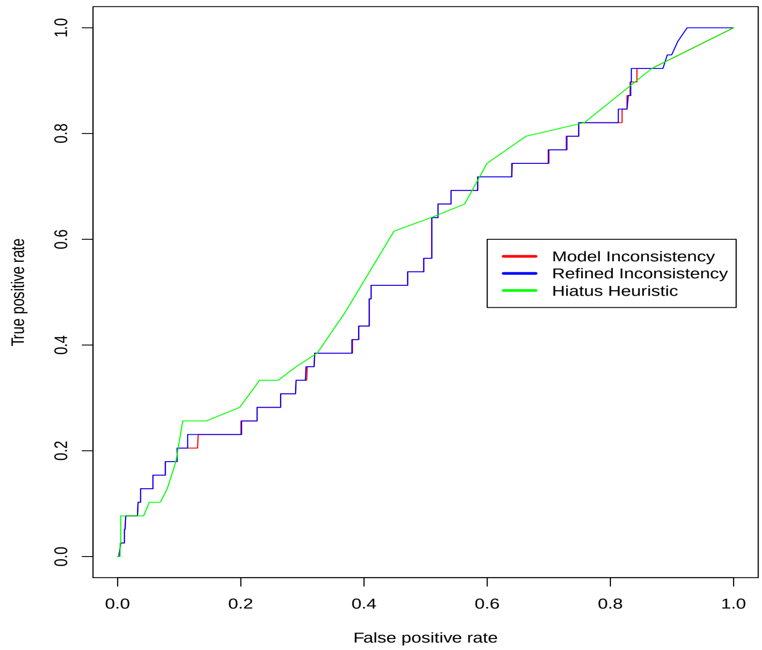

Figure 6.6: DMPT/Apparel ROC Analysis of Predicting Customer Inactivity with Model Inconsistency and Hiatus Heuristic

retail datasets. The refined rule (blue curve) dominates the original rule (red curve) on all three datasets, even though the differences are small. The intersections between the green curve (hiatus heuristic) and the blue curve indicate that neither method dominates the other completely on any of the datasets. It should be noted that the ROC curves in Figure 6.6 suggest very poor classification accuracy for the apparel dataset by all methods. All three curves are near the diagonal, which signifies a near chance result.

The visual impression of the ROC curves is corroborated by the corresponding AUC measures compiled in Table 6.4. The refined rule produces the highest AUC for the DIY data (0.8816) and CDNOW data (71.11), while the hiatus heuristic performs better on the apparel data (0.5879). The original rule based on model inconsistency generates the lowest AUC for all three datasets.

I additionally calculated the forecasting accuracy at the "natural" cutoff of 95% certainty for the DMPT (refined rule) and the commonly used four months hiatus cutoff for the heuristic. The heuristic is then simply:

> *If a customer has not made a purchase for at least four months, the customer is classified as inactive; otherwise the customer is classified as active.*

The identification of individuals as either active or inactive leads to four possible classification results. The classification of an individual customer can be correct or incorrect and the customer can be classified as active or inactive. The results of the heuristic and the DMPT (refined rule) are summarized in Table 6.5. At the "natural" cutoff values the heuristic has a higher overall classification accuracy for the CDNOW data (64.72% vs. 61.16%) and the apparel data (87.03% vs. 85.87%), whereas the DMPT (refined rule) classifies marginally more customers

Table 6.4: ROC/AUC Analysis of Predicting Customer Inactivity

	DIY Retailer	**Apparel Retailer**	**CDNOW**
Hiatus Heuristic:	0.8762	0.5879	0.7099
DMPT Model Inconsistency:	0.8704	0.5602	0.6960
DMPT Refined Inconsistency:	0.8816	0.5634	0.7111

Table 6.5: DMPT - Accuracy of Predicting Customer Inactivity (%)

	CDNOW		Apparel Retailer		DIY Retailer	
	DMPT	**Heuristic**	**DMPT**	**Heuristic**	**DMPT**	**Heuristic**
True Positive Rate:	32.24	42.11	23.08	25.64	45.45	40.91
True Negative Rate:	89.17	86.62	88.33	89.44	96.91	96.91
False Negative Rate:	67.76	57.89	76.92	74.36	54.55	59.09
False Positive Rate:	10.83	13.38	11.67	10.56	3.09	3.09
Overall Error:	38.83	35.28	14.13	12.97	3.95	4.03
Overall Accuracy:	61.17	64.72	85.87	87.03	96.05	95.97

of the DIY retailer correctly (96.05% vs. 95.97%).

In regard to the correctly classified active customers (true negative rate) the results are extremely close: the accuracy for the DIY retailer is dead even (96.91% both), the DMPT is ahead on the CDNOW dataset (89.17% vs. 86.62%) and the heuristic is better on the apparel retail data (89.44% vs. 88.33%).

6.3.2.2 Long- and Short-Term Predictions

The long- and short-term forecasts generated by the DMPT and the heuristic indicate that neither method dominates the other. One pattern that emerges is that the DMPT is, in general, more accurate on the median aggregated metrics and on an overall level *MAPE*, while the heuristic is more accurate according to outlier sensitive mean aggregated metrics.

In this context long-term prediction of future interpurchase times refers to a forecast horizon covering the full length of the hold-out period. Short-term prediction entails the forecast of the next single interpurchase time.

The baseline model is the heuristic:

Customers continue to buy at their mean past purchase frequency.

Table 6.6 contains the short-term forecast results for all three retail datasets. The overall (non-individual) level accuracy (*MAPE*) of the DMPT is higher than the accuracy of the managerial heuristic on all three datasets. The differences in accuracy range from 2.69% for the DIY retailer to 12.45% for CDNOW. The results for

Table 6.6: DMPT - Accuracy of Short-Term Interpurchase Time Prediction

	CDNOW		Apparel Retailer		DIY Retailer	
	DMPT	**Heuristic**	**DMPT**	**Heuristic**	**DMPT**	**Heuristic**
MAPE (%)	76.26	88.71	67.36	72.45	29.63	32.32
Mean MSE	126.338	123.094	201.393	184.865	2.549	2.577
Median MSE	24.289	32.111	33.358	40.111	0.060	0.111
Mean RMSE	7.786	7.741	9.844	9.508	0.699	0.701
Median RMSE	4.928	5.667	5.775	6.333	0.245	0.333

Table 6.7: DMPT - Accuracy of Long-Term Interpurchase Time Prediction

	CDNOW		Apparel Retailer		DIY Retailer	
	DMPT	**Heuristic**	**DMPT**	**Heuristic**	**DMPT**	**Heuristic**
MAPE (%)	112.88	117.88	137.26	150.61	37.84	33.93
Mean MSE	111.860	107.405	133.136	126.191	1.943	1.834
Median MSE	25.071	36.000	34.511	41.143	0.256	0.222
Mean RMSE	7.781	7.537	8.416	8.452	0.818	0.753
Median RMSE	5.007	6.000	5.874	6.414	0.506	0.471

the aggregated individual level error metrics are mixed. The less outlier sensitive metrics *median MSE* and *median RMSE* favor the DMPT on all three datasets. Judged by the more outlier sensitive mean aggregated metrics the heuristic is more accurate for the CDNOW and apparel data.

The long-term forecasts are shown in Table 6.7. The DMPT performs better than the heuristic by *MAPE* for the customers of CDNOW (112.88 vs. 117.88) and the apparel retailer (137.26 vs. 150.61). For these datasets the DMPT also outperforms the heuristic on the less outlier sensitive error metrics *median MSE* and *median RMSE*. The results are reversed for the DIY data, on which the heuristic is more accurate. The heuristic is also more accurate according to the outlier-sensitive metrics for CDNOW (*mean MSE*: 107.405 vs. 111.860) and the apparel dataset (*mean MSE*: 126.191 vs. 133.136). The only exception is the *mean RMSE* on the apparel data, where the DMPT is marginally better (8.416 vs. 8.452).

6.3.2.3 Prediction of Future Best Customers

The heuristic outperforms the DMPT model consistently on all three datasets in predicting the top 10% and 20% of customers. The DMPT is tested by rank-ordering customers according to the model's forecast of purchases frequencies for the hold-out period. The actual future top 10% (20%) of customers are determined by their observed purchase frequency in the hold-out period. I benchmarked the DMPT against the following baseline heuristic:

> *The past 10% (20%) best customers will also be the future 10% (20%) best customers.*

Tables 6.8 and 6.9 present the results for predicting the top 10% and 20% best customers of each customer base. The DMPT exhibits the lowest classification rate when predicting the future top 10% of customers in the CDNOW dataset (3.33%).

Table 6.8: DMPT - Accuracy of Future Top 10% Customer Prediction (%)

	CDNOW		**Apparel Retailer**		**DIY Retailer**	
	DMPT	**Heuristic**	**DMPT**	**Heuristic**	**DMPT**	**Heuristic**
True Positive Rate:	3.33	46.67	50.49	55.34	16.79	51.91
True Negative Rate:	89.61	94.27	94.52	95.05	90.80	94.68
False Negative Rate:	96.67	53.33	49.51	44.66	83.21	48.09
False Positive Rate:	10.39	5.73	5.48	4.95	9.20	5.32
Overall Error:	18.77	10.36	9.87	8.91	16.57	9.57
Overall Accuracy:	81.23	89.64	90.13	91.09	83.43	90.43

Table 6.9: DMPT - Accuracy of Future Top 20% Customer Prediction (%)

	CDNOW		**Apparel Retailer**		**DIY Retailer**	
	DMPT	**Heuristic**	**DMPT**	**Heuristic**	**DMPT**	**Heuristic**
True Positive Rate:	8.20	52.46	49.02	58.82	15.59	51.71
True Negative Rate:	77.42	88.31	97.35	97.86	78.92	87.94
False Negative Rate:	91.80	47.54	50.98	41.18	84.41	48.29
False Positive Rate:	22.58	11.69	2.65	2.14	21.08	12.06
Overall Error:	36.25	18.77	5.03	4.07	33.74	19.30
Overall Accuracy:	63.75	81.23	94.97	95.93	66.26	80.70

6.4 Discussion of Results

The DMPT performs relatively best when predicting long- and short-term interpurchase times. In this respect, the DMPT is slightly more accurate than the heuristic on an overall level and on most non-outlier-sensitive individual level metrics. The heuristic is more accurate than the DMPT according to some of the outlier sensitive error measures. This indicates that for the majority of customers the DMPT yields better results than the heuristic, but also that there are some customers for which the DMPT produces relatively large forecast errors. Nonetheless, both methods are very close in forecast accuracy without one being clearly better than the other.

My refined rule for deriving the probability of individual customer inactivity, which takes into account the actual predictive distribution and the finite length of the hold-out period, yields a slight but consistent improvement on previous approaches for the DMPT. The ROC analysis on three large datasets shows that the new scheme is equal or better over the complete range of potential cutoff values. However, compared to the hiatus heuristic the results of the DMPT, even with the refined rule, are mixed. The refined rule performs better for the DIY and CDNOW data, while the hiatus heuristic is better for the apparel data according to AUC. Neither scheme dominates the other over the full parameter space in the ROC analysis.

It should be noted that all methods struggled with the customer base of the apparel retailer. The prediction of customer inactivity for this dataset did not generate results distinctly different from random guessing. One reason might be the limitation of the DMPT to customers with at least four observed interpurchase times. The restriction to this subset of customers could have distorted the classification results.

The relatively worse performance of the DMPT in rank-ordering customers was to be expected as the model is not designed for this task and similar results have been reported by Huang (2012) and Wübben and von Wangenheim (2008) for the Pareto/NBD and BG/NBD model. It is still surprising how well the simple heuristics performs here as it produces more accurate forecasts on all three datasets.

As, simple heuristics can perform rather well on "big data", the decision to implement more complex methods should be taken with great care. It seems that the temporal dynamic part of the DMPT (which captures changes in purchase trends) and the component mixture (which assumes distinct groups of customers) adds only little information that improves out-of-sample prediction on the datasets used in the comparison.

The mixed results of the DMPT are the motivation for developing two new hierarchical Bayesian models that take into account seasonal factors (HSM) and include an explicit drop-out process (HSMDO).

Chapter 7

Hierarchical Bayesian Seasonal Effects Model

In this chapter I develop the hierarchical Bayesian seasonal effects model (HSM). The model predicts future purchase levels and operates under the "always a share" assumption. The model aims at (1) improving predictive accuracy by incorporating seasonal effects, (2) capturing individual customers' seasonal behavior in relation to group-level seasonal behavior, and at (3) relating individual purchase rates and cross-sectional heterogeneity. The model yields a measure of individual seasonality, that indicates how strongly the customer follows the cross-sectional seasonality, whether he purchases anti-seasonal, or if the customer's behavior is non-seasonal. In business practice such a measure is particularly useful for purposes of targeting groups of customers, customer segmentation and timing of marketing actions.

The "always a share" assumption implies that the customer is always alive and does not defect, but still might exhibit long periods with few or no purchases. Similar to the analysis of the DMPT, I use three retail datasets for parameter estimation and model comparison. I assess the predictive performance of the model in relation to baseline managerial heuristics, Holt-Winters method, SARIMA models and discuss the results based on this comparison.

7.1 General Model Framework

This model contributes to the literature stream of probability models for customer base analysis and follows an evolutionary model-building view (Fader and Hardie 2009). The HSM is based on a discretely sampled inhomogeneous Poisson counting process and is unique in that it combines the following features:

1. a hierarchical purchase process that relates the distribution of individual level purchase frequency over time to cross-sectional heterogeneity;
2. a hierarchical seasonal structure that relates individual level seasonality to cross-sectional seasonal components in a multiplicative submodel; and
3. an individual estimate for each customer's seasonality that indicates how strongly he follows the cross-sectional seasonality, whether he purchases anti-seasonal, or if he exhibits non-seasonal behavior.

The hierarchical structure of model that captures individual purchase rates and cross-sectional heterogeneity is closely related to previous work in this research stream such as the compound Poisson model (Ehrenberg 1959), the Pareto/NBD model (Schmittlein, Morrison, and Colombo 1987), the BG/NBD model (Fader, Hardie, and Lee 2005a), the PDO/NBD model (Jerath, Fader, and Hardie 2011), the G/G/NBD model (Bemmaor and Glady 2012), the BG/GCP model (Mzoughia and Limam 2014), and related variants (Glady, Lemmens, and Croux 2015). However, as of yet no attempt to include individual and cross-sectional seasonal effects in the Pareto/NBD model framework have been made.

The seasonal submodel of the HSM generalizes previous post-hoc ratio to moving average adjustments of forecasts obtained with the Pareto/NBD model (Zitzlsperger, Robbert, and Roth 2009) and the inclusion of dummy variables (Schweidel and Knox 2013). The models in this thesis aim at closing the gap Ballings and Van den Poel (2015, p. 257) point out when they remark that "taking seasonality into account would only increase the predictive performance of our models."

In the following I present the HSM in three parts that build upon each other and together form the complete model:

1. the Poisson covariate gamma mixture;
2. the hierarchical seasonal effects submodel; and
3. prior distributions, hyperparameters and parameter space.

7.1.1 Poisson Covariate Gamma Mixture

Each customer i with $i \in \{1..N\}$ is assumed to buy at an individual rate $\lambda_i e^{\beta_i s_{k(j)}}$ governed by a Poisson purchase process[14]. Customers are not defecting under an "always a share" assumption. An individual customer i's purchases at time j are denoted $x_{ij} \in \mathbb{Z}_{\geq 0}$.

The model extends the NBD Model (Ehrenberg 1988, p. 128) structure, in which the expected long-run purchase for customer i per time unit is λ_i, by including the multiplicative random-effects submodel $e^{\beta_i s_{k(j)}}$ as a time-varying co-variate. The heterogeneity across the customers' individual purchase rates λ_i is captured by a gamma distribution. The parameters of the group level gamma distribution are the scale parameter $r > 0$ and the shape parameter $\alpha > 0$:

$$x_{ij} \sim \quad Poisson(\lambda_i e^{\beta_i s_{k(j)}}) \stackrel{\text{p.m.f.}}{\rightleftharpoons} e^{-\lambda_i e^{\beta_i s_{k(j)}}} \frac{(\lambda_i e^{\beta_i s_{k(j)}})^{x_{ij}}}{\Gamma(x_{ij}+1)} \tag{7.1}$$

$$\lambda_i \sim \quad Gamma(r, \alpha) \stackrel{\text{p.d.f.}}{\rightleftharpoons} e^{-\lambda_i \alpha} \lambda_i^{r-1} \frac{\alpha^r}{\Gamma(r)} \tag{7.2}$$

The time-varying random-effects submodel $\beta_i s_{k(j)}$ describes the seasonal structure of the model. The term depends both on time j and customer i and reflects both the joint seasonal structure on a group level as well as individual level seasonal effects as explained below.

[14] This process can be seen as an discrete sampled inhomogeneous Poisson counting process with $N(j) - N(j-1) = \Delta N_j \sim Poisson(\int_{j-1}^{j} \lambda(t)dt)$, where $\lambda(t)$ is constant between $(j-1, j]$ and therefore $\Delta N_j \sim Poisson(\lambda(j))$.

7.1.2 Hierarchical Seasonal Effects Submodel

The individual level seasonality of each customer i is captured by a multiplicative random effect β_i. It reflects the influence of the seasonal components $s_{k(j)}$ on each customers individual purchase rate λ_i. This results in the Poisson purchase rate $\lambda_i e^{\beta_i s_{k(j)}}$ for customer i in time period j. Cross sectional heterogeneity in β_i is assumed to be distributed normal with mean one and precision[15] τ_β.

I constructed the model to comprehensively capture seasonal behavior. For example, if an individual customer i exhibits amplified (pro)seasonal behavior his seasonality coefficient would be $\beta_i > 1$. If a customer shows an attenuated seasonal pattern, compared to the cross-sectional seasonal components, then his coefficient would be $0 < \beta_i < 1$. If a customer purchases fairly constantly at his individual purchase rate λ_i, implying non-seasonal behavior, his coefficient β_i would be near or at zero. Finally, if a customer shows anti-seasonal purchase patterns, behaving opposite to the overall seasonal behavior, his coefficient would be $\beta_i < 0$.

Let s_k denote the k^{th} seasonal component with $k \in \{1..K\}$. Each point in time j is linked[16] to one of the K seasonal components by $k(j) = ((j-1) \bmod K) + 1$. The (uncentered) seasonal components $\hat{s}_k$ are assumed to be exchangeable and distributed normal with mean zero and precision τ_s. The components $\hat{s}_k$ are centered at zero (Equation 7.4), so that s_k reflect the (approximate) seasonal percentage change compared to average purchases levels.

The structure of the seasonal submodel can then be decomposed and summarized

[15] In Bayesian modeling normal distributions are often specified with the precision parameter $\tau = 1/\sigma^2$ instead of the variance σ^2.

[16] The model allows for any form of time-varying multiplicative co-variates and is derived in full generality. For the general case the link function may be set to the identity function $k(j) = j$

as follows:

$$\hat{s}_k \sim \quad Normal(0, \tau_s) \quad \overset{\text{p.d.f.}}{\rightleftharpoons} \sqrt{\frac{\tau_s}{2\pi}}\, e^{-\tau_s \hat{s}_k^2/2} \tag{7.3}$$

$$s_k = \quad \hat{s}_k - \tfrac{1}{K}\sum \hat{s}_k \tag{7.4}$$

$$\beta_i \sim \quad Normal(1, \tau_\beta) \quad \overset{\text{p.d.f.}}{\rightleftharpoons} \sqrt{\frac{\tau_\beta}{2\pi}}\, e^{-\tau_\beta (\beta_i - 1)^2/2} \tag{7.5}$$

$$k(j) = \quad ((j-1) \bmod K) + 1 \tag{7.6}$$

In a Bayesian hierarchical model one can include all K seasonal components s_k and N seasonal parameters β_i as they are drawn from distributions with finite precisions and that group level information allows for identification. The parameters of the seasonal model are $\tau_s > 0$ and $\tau_\beta > 0$.

7.1.3 Priors, Hyperparameters and Parameter Space

The hyperparameters of the prior distributions are chosen so that the precision parameters (τ_s, τ_β) and the parameters of the Poisson-gamma mixture (μ_λ, σ_λ) are given vague non-informative gamma priors ($\varepsilon = 0.001$). The parameters r and α are re-parameterized in terms of a mean parameter μ_λ and a scale parameter σ_λ. Together with the following transformations these distributions complete the model:

$$r = \mu_\lambda^2/\sigma_\lambda \quad \alpha = \mu_\lambda/\sigma_\lambda \tag{7.7}$$

$$\sigma_\lambda, \mu_\lambda, \tau_s, \tau_\beta \sim \quad Gamma(\varepsilon, \varepsilon) \tag{7.8}$$

Let N be the number of customers and K the number of seasonal components then the model contains $2N+K+4$ parameters of interest. The term $2N+K+4$ results from N individual purchase frequencies parameters λ_i, N individual seasonality measures β_i, K seasonal components, and four parameters $\sigma_\lambda, \mu_\lambda, \tau_s$ and τ_β, while r and α are just transformations of $\sigma_\lambda, \mu_\lambda$.

7.1.4 Derivation of Model Likelihood

First, I derive the individual level likelihood function. For better readability I set $s_{k(j)} = s_j$ and let s denote the set of s_k with $k \in \{1..K\}$. Then conditional on $(X_{ij} = x_{ij},\ T,\ s)$ and using Equation (7.1), the individual level likelihood for customer i is:

$$L_i(\lambda_i, \beta_i | X_{ij} = x_{ij}, T, s) = \prod_{j=1}^{T} Poisson(x_{ij} | \lambda_i e^{\beta_i s_j}) \tag{7.9}$$

In order to derive the sample-likelihood function that includes all customers, let λ and β be vectors over all customers with, e.g., λ_i denoting the i^{th} element of the vector. Using Bayes' Theorem, the prior definitions (7.2), (7.3) and (7.5) and aggregating over all customers yields the sample-likelihood:

$$\begin{aligned} &L(\lambda, \beta, r, \alpha, \hat{s}, \tau_\beta, \tau_s | X, T) = \\ &\quad \Big[\prod_{i=1}^{N} L_i(\lambda_i, \beta_i | X_{ij} = x_{ij}, T, s_k = \hat{s}_k - \frac{1}{K}\sum \hat{s}_k)\Big] \Big[\prod_{k=1}^{K} Normal(\hat{s}_k | 0, \tau_s)\Big] \\ &\quad \Big[\prod_{i=1}^{N} Normal(\beta_i | 1, \tau_\beta) Gamma(\lambda_i | r, \alpha)\Big] \end{aligned} \tag{7.10}$$

The joint posterior distribution results by substituting r and α (7.7), and multiplying the sample likelihood (7.10) with the priors (7.8):

$$\begin{aligned} &p(\lambda, \beta, r, \alpha, \hat{s}, \tau_\beta, \tau_s | X, T) \propto \\ &\quad L(\lambda, \beta, r = \mu_\lambda^2/\sigma_\lambda, \alpha = \mu_\lambda/\sigma_\lambda, \hat{s}, \tau_\beta, \tau_s | X, T) Gamma(\tau_\beta | \varepsilon, \varepsilon) \\ &\quad Gamma(\tau_s | \varepsilon, \varepsilon) Gamma(\sigma_\lambda | \varepsilon, \varepsilon) Gamma(\mu_\lambda | \varepsilon, \varepsilon) \end{aligned} \tag{7.11}$$

7.2 Data Analysis: Parameter Estimation and Prediction

7.2.1 Convergence and Marginal Posterior Estimates

The training of the HSM consists of learning about the main parameters posterior distributions. For training, I only use data from the calibration period of each retailer and reserve the hold-out data for assessing the predictive performance and for model comparison.

I obtained the posterior distributions for the model's parameters by a MCMC procedure via Gibbs sampling. The source code and additional information about the procedure are presented in Appendix B.1. The parameter estimates are based on MCMC runs with four chains starting from dispersed initial values, 20,000 iterations and a burn-in phase of 10,000 samples. I performed thinning by taking every fifth sample after the burn-in period to save memory. Therefore, the posterior distributions are based on $(10,000/5) \times 4 = 8,000$ samples.

The MCMC procedure iteratively converges towards the true posterior distribution of the parameters and after convergence simulates their true posterior distribution. I assessed approximate convergence of the algorithm by examining auto-correlation and effective sample size; visually inspecting the trace plots of the posterior samples; evaluating the Brooks, Gelman and Rubin statistic (BGR); and by running multiple simulations from dispersed initial values.

Auto-Correlation and Effective Sample Size

The samples obtained for the CDNOW, DIY and apparel retailers show virtually no or very little auto-correlation. The auto-correlation plots for the main parameters of the model and seasonal components are shown in Figures B.1 and B.3 (Apparel), Figures B.5 and B.7 (DIY), and Figures B.9 and B.11 (CDNOW) in the Appendix.

The lowest two effective sample sizes for the DIY MCMC simulation are 1,305 for τ_β and 1,752 for s_{10}, while the two lowest effective sample sizes for the apparel samples are 2,744 for s_1 and 2,762 for τ_β. On the CDNOW dataset the fewest effective samples were collected for α (3,435) and τ_β (4,455). Typically, effective sample sizes were near their theoretical maximum of 8,000 samples. Thus, autocorrelation is not an issue with the obtained effective sample sizes, even if τ_β's mixing is slightly worse than that of the other parameters (Gelman et al. 2003, p. 298).

Inspection of Trace Plots

The trace plots for all three datasets exhibit no visible "trends" and sufficient mixing of all four Markov chains after a burn-in period of 10,000 iterations. The trace plots for the main parameters of the model and all seasonal components are shown in Figures B.2 and B.4 (Apparel), Figures B.6 and B.8 (DIY), and Figures B.10 and B.12 (CDNOW) in the Appendix. The visual inspection of the trace plots indicates that convergence was reached and that the posterior distribution is adequately represented (see Section 4.3).

BGR Statistic

I validated Markov chain convergence by computing the BGR statistic for all parameters over all four chains. In all cases, the BGR ratio $\hat{R}_c$ stayed well below 1.1 close to the ideal 1.0, which indicates that the chains have converged to the true posterior distribution (Congdon 2003; Cowles and Carlin 1996; Gelman and Rubin 1992). In addition, I repeatedly simulated the posterior distribution using multi-chain runs with dispersed initial values which in all cases converged to virtually identical parameter values.

Parameter Estimates and Posterior Densities

Summary statistics for the model's parameters including means, standard deviations, medians, and credibility intervals are compiled in Tables 7.1 (DIY), 7.2 (Apparel), and 7.3 (CDNOW). The tables include all main model parameters r, α, τ_s, and τ_β; all seasonal components s_k; and the first and last two individual level parameters, β_i and λ_i. The resulting full marginal posterior densities for the main parameters and the seasonal components are depicted in Figures 7.4-7.9.

Table 7.1: HSM/DIY Marginal Posterior Distribution Summaries for MCMC Simulation after 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

Summary of Marginal Posterior Distributions						
node	*mean*	*sd*	*MC error*	*2.5%*	*median*	*97.5%*
r	2.1466	0.0801	0.0009	1.9934	2.1458	2.3067
alpha	0.7719	0.0322	0.0004	0.7105	0.7718	0.8366
tau.s	75.0392	33.6800	0.3766	25.0396	70.0170	156.4308
tau.beta	0.4357	0.0578	0.0007	0.3299	0.4332	0.5556
s[1]	0.0646	0.0076	0.0001	0.0499	0.0646	0.0795
s[2]	0.0356	0.0082	0.0001	0.0194	0.0356	0.0515
s[3]	0.0193	0.0075	0.0001	0.0048	0.0193	0.0340
s[4]	−0.0028	0.0094	0.0001	−0.0215	−0.0028	0.0156
s[5]	−0.0140	0.0104	0.0001	−0.0347	−0.0140	0.0059
s[6]	−0.0941	0.0111	0.0001	−0.1165	−0.0940	−0.0731
s[7]	−0.2317	0.0170	0.0002	−0.2655	−0.2317	−0.1986
s[8]	−0.1710	0.0147	0.0002	−0.2005	−0.1707	−0.1430
s[9]	0.0360	0.0091	0.0001	0.0181	0.0360	0.0543
s[10]	0.1448	0.0106	0.0001	0.1242	0.1447	0.1656
s[11]	0.1270	0.0093	0.0001	0.1086	0.1269	0.1451
s[12]	0.0864	0.0086	0.0001	0.0698	0.0863	0.1034
lambda[1]	7.5565	0.5232	0.0059	6.5582	7.5444	8.6004
lambda[2]	2.6938	0.3145	0.0035	2.1105	2.6825	3.3362
⋮	⋮	⋮	⋮	⋮	⋮	⋮
lambda[1328]	1.3800	0.2230	0.0025	0.9773	1.3683	1.8419
lambda[1329]	4.4143	0.3961	0.0044	3.6683	4.4034	5.2140
beta[1]	2.6519	0.6927	0.0077	1.3083	2.6499	4.0586
beta[2]	2.0284	0.9625	0.0108	0.2079	2.0174	4.0112
⋮	⋮	⋮	⋮	⋮	⋮	⋮
beta[1328]	1.6031	1.1380	0.0127	−0.5582	1.5893	3.8988
beta[1329]	0.8482	0.7785	0.0087	−0.6259	0.8337	2.4201

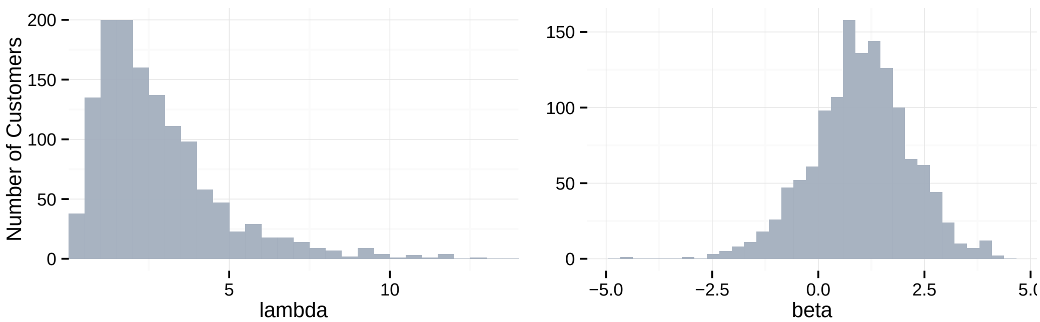

Figure 7.1: HSM/DIY Customer Base Histograms for Estimates of Individual Customer Seasonality β_i and Purchase Rate λ_i

Table 7.2: HSM/Apparel Marginal Posterior Distribution Summaries for MCMC Simulation after 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

Summary of Marginal Posterior Distributions						
node	*mean*	*sd*	*MC error*	*2.5%*	*median*	*97.5%*
r	1.1155	0.0456	0.0005	1.0274	1.1147	1.2071
alpha	1.7920	0.0884	0.0010	1.6237	1.7896	1.9662
tau.s	5.5811	2.4178	0.0270	1.8773	5.2332	11.2034
tau.beta	1.5209	0.1423	0.0016	1.2631	1.5174	1.8152
s[1]	0.8552	0.0291	0.0003	0.7992	0.8550	0.9118
s[2]	−0.4428	0.0560	0.0006	−0.5561	−0.4420	−0.3365
s[3]	−0.1999	0.0421	0.0005	−0.2847	−0.1990	−0.1214
s[4]	−0.1129	0.0442	0.0005	−0.2035	−0.1115	−0.0298
s[5]	−0.1183	0.0423	0.0005	−0.2028	−0.1170	−0.0370
s[6]	0.0499	0.0324	0.0004	−0.0150	0.0501	0.1117
s[7]	0.2530	0.0287	0.0003	0.1968	0.2529	0.3096
s[8]	0.0361	0.0364	0.0004	−0.0366	0.0373	0.1038
s[9]	−0.7434	0.0937	0.0011	−0.9323	−0.7398	−0.5691
s[10]	0.1298	0.0432	0.0005	0.0435	0.1313	0.2112
s[11]	−0.2229	0.0569	0.0006	−0.3395	−0.2212	−0.1171
s[12]	0.5162	0.0316	0.0004	0.4546	0.5157	0.5781
lambda[1]	1.6030	0.2762	0.0031	1.1176	1.5849	2.1942
lambda[2]	2.3291	0.3298	0.0037	1.7221	2.3123	3.0150
⋮	⋮	⋮	⋮	⋮	⋮	⋮
lambda[1229]	0.5138	0.1515	0.0017	0.2625	0.4973	0.8477
lambda[1230]	1.2585	0.2373	0.0027	0.8271	1.2455	1.7630
beta[1]	0.4806	0.3756	0.0042	−0.2615	0.4820	1.2027
beta[2]	0.6415	0.3228	0.0036	−0.0060	0.6485	1.2495
⋮	⋮	⋮	⋮	⋮	⋮	⋮
beta[1229]	0.8126	0.5390	0.0060	−0.2575	0.8187	1.8720
beta[1230]	0.6974	0.4068	0.0046	−0.1082	0.6987	1.5044

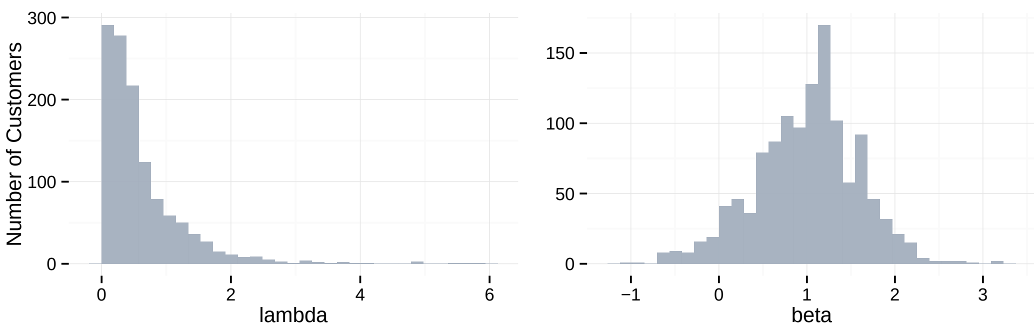

Figure 7.2: HSM/Apparel Customer Base Histograms for Estimates of Individual Customer Seasonality β_i and Purchase Rate λ_i

Table 7.3: HSM/CDNOW Marginal Posterior Distribution Summaries for MCMC Simulation after 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

Summary of Marginal Posterior Distributions

node	*mean*	*sd*	*MC error*	*2.5%*	*median*	*97.5%*
r	1.9758	0.1573	0.0018	1.6955	1.9685	2.3011
alpha	15.1102	1.3336	0.0149	12.6979	15.0641	17.8570
tau.s	3.5915	1.5494	0.0173	1.2355	3.3746	7.1694
tau.beta	811.2189	692.2109	7.7392	130.7259	595.2211	2813.8055
s[1]	1.5192	0.0421	0.0005	1.4352	1.5199	1.6004
s[2]	0.1398	0.0664	0.0007	0.0091	0.1400	0.2658
s[3]	0.2374	0.0847	0.0010	0.0704	0.2377	0.4003
s[4]	−0.0270	0.0932	0.0010	−0.2158	−0.0244	0.1511
s[5]	−0.1172	0.0990	0.0011	−0.3200	−0.1149	0.0696
s[6]	−0.0948	0.0985	0.0011	−0.2946	−0.0942	0.0950
s[7]	−0.0962	0.0967	0.0011	−0.2910	−0.0951	0.0883
s[8]	−0.3143	0.1092	0.0012	−0.5325	−0.3130	−0.1056
s[9]	−0.2770	0.1059	0.0012	−0.4924	−0.2752	−0.0775
s[10]	−0.4787	0.1161	0.0013	−0.7114	−0.4770	−0.2585
s[11]	−0.1923	0.1013	0.0011	−0.3977	−0.1904	−0.0033
s[12]	−0.2990	0.1078	0.0012	−0.5159	−0.2982	−0.0961
lambda[1]	0.1702	0.0694	0.0008	0.0637	0.1606	0.3274
lambda[2]	0.1119	0.0564	0.0006	0.0303	0.1025	0.2461
⋮	⋮	⋮	⋮	⋮	⋮	⋮
lambda[780]	0.0838	0.0484	0.0005	0.0175	0.0754	0.2033
lambda[781]	0.0848	0.0489	0.0006	0.0166	0.0759	0.2035
beta[1]	0.9998	0.0475	0.0005	0.9020	0.9996	1.0970
beta[2]	1.0034	0.0476	0.0005	0.9056	1.0030	1.1044
⋮	⋮	⋮	⋮	⋮	⋮	⋮
beta[780]	1.0007	0.0479	0.0005	0.8991	1.0007	1.0972
beta[781]	1.0007	0.0481	0.0005	0.9024	1.0003	1.0993

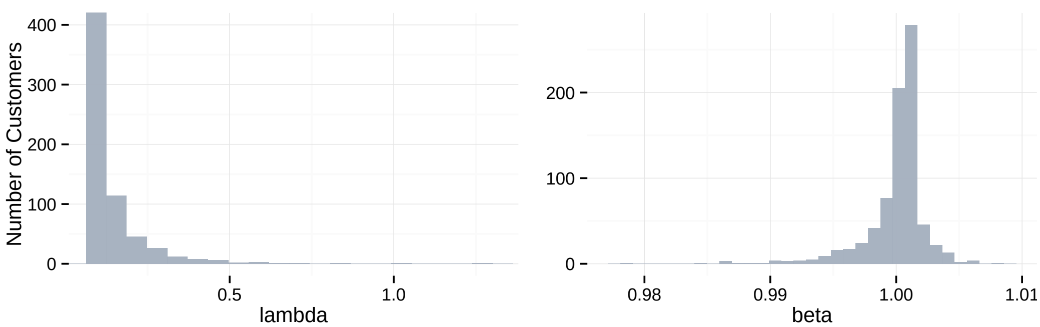

Figure 7.3: HSM/CDNOW Customer Base Histograms for Estimates of Individual Customer Seasonality β_i and Purchase Rate λ_i

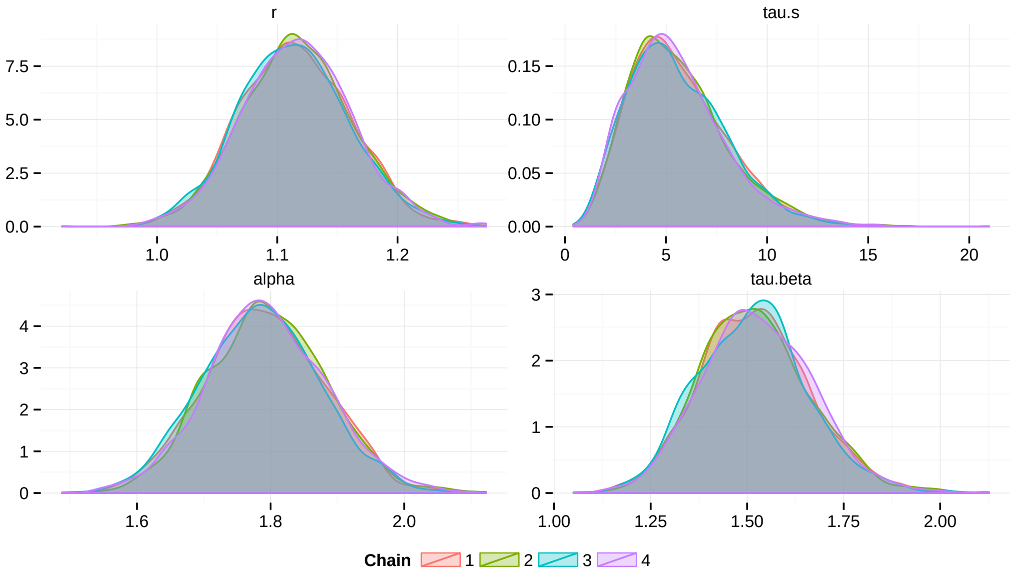

Figure 7.4: HSM/Apparel Marginal Posterior Densities for r, α, τ_s, and τ_β after 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

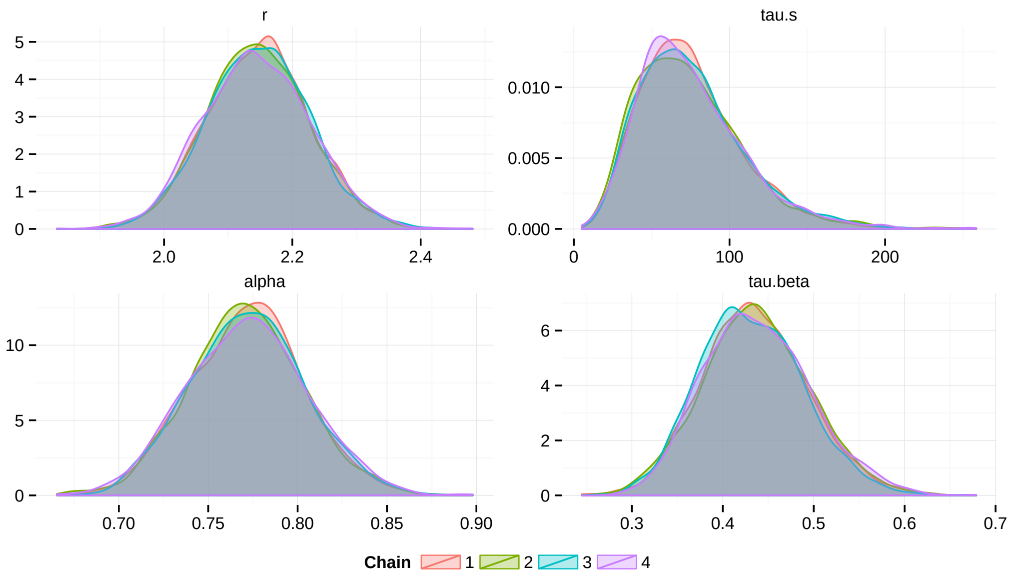

Figure 7.5: HSM/DIY Marginal Posterior Densities for r, α, τ_s, and τ_β after 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

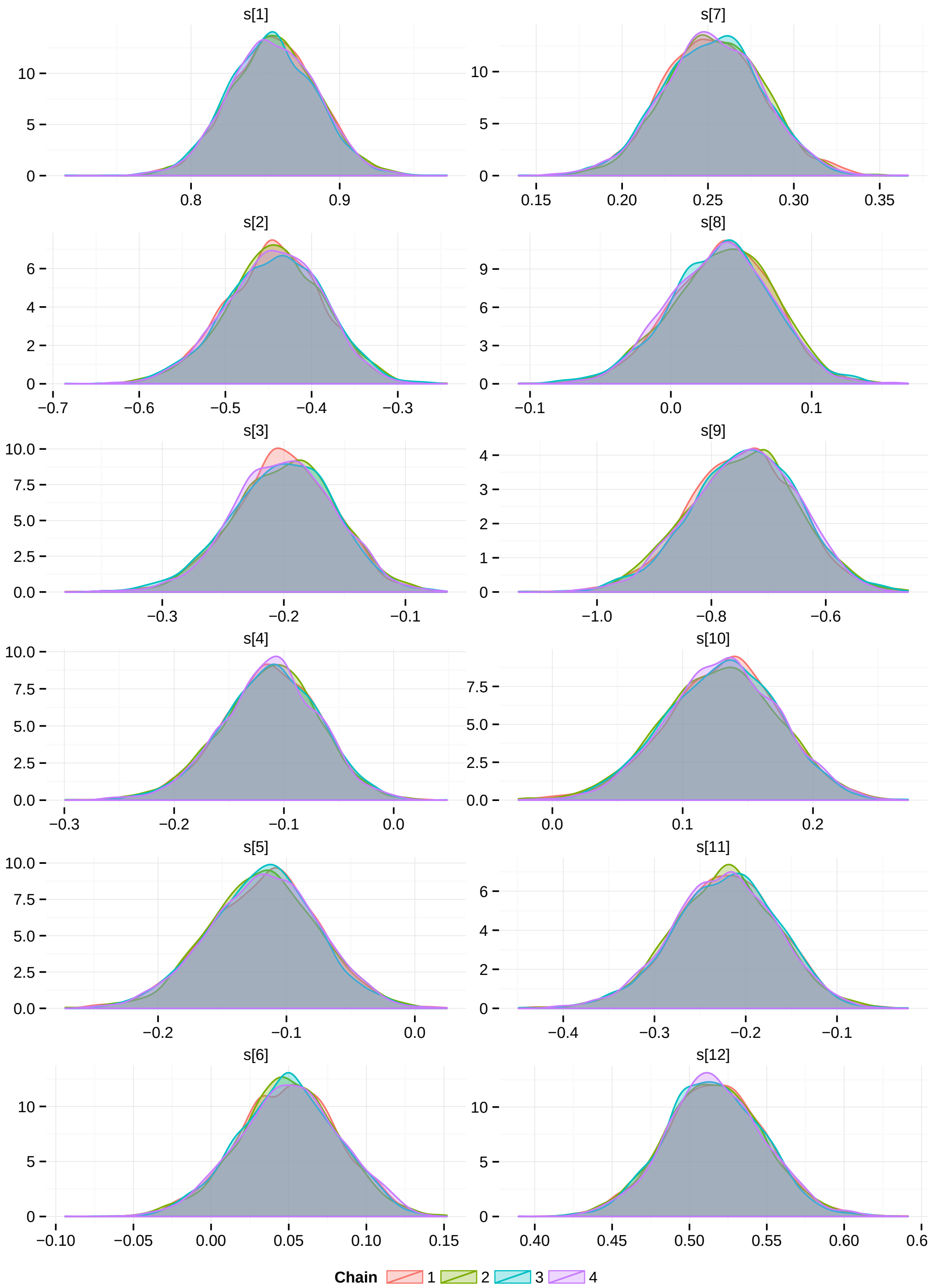

Figure 7.6: HSM/Apparel Marginal Posterior Densities for Seasonal Components s_k after 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

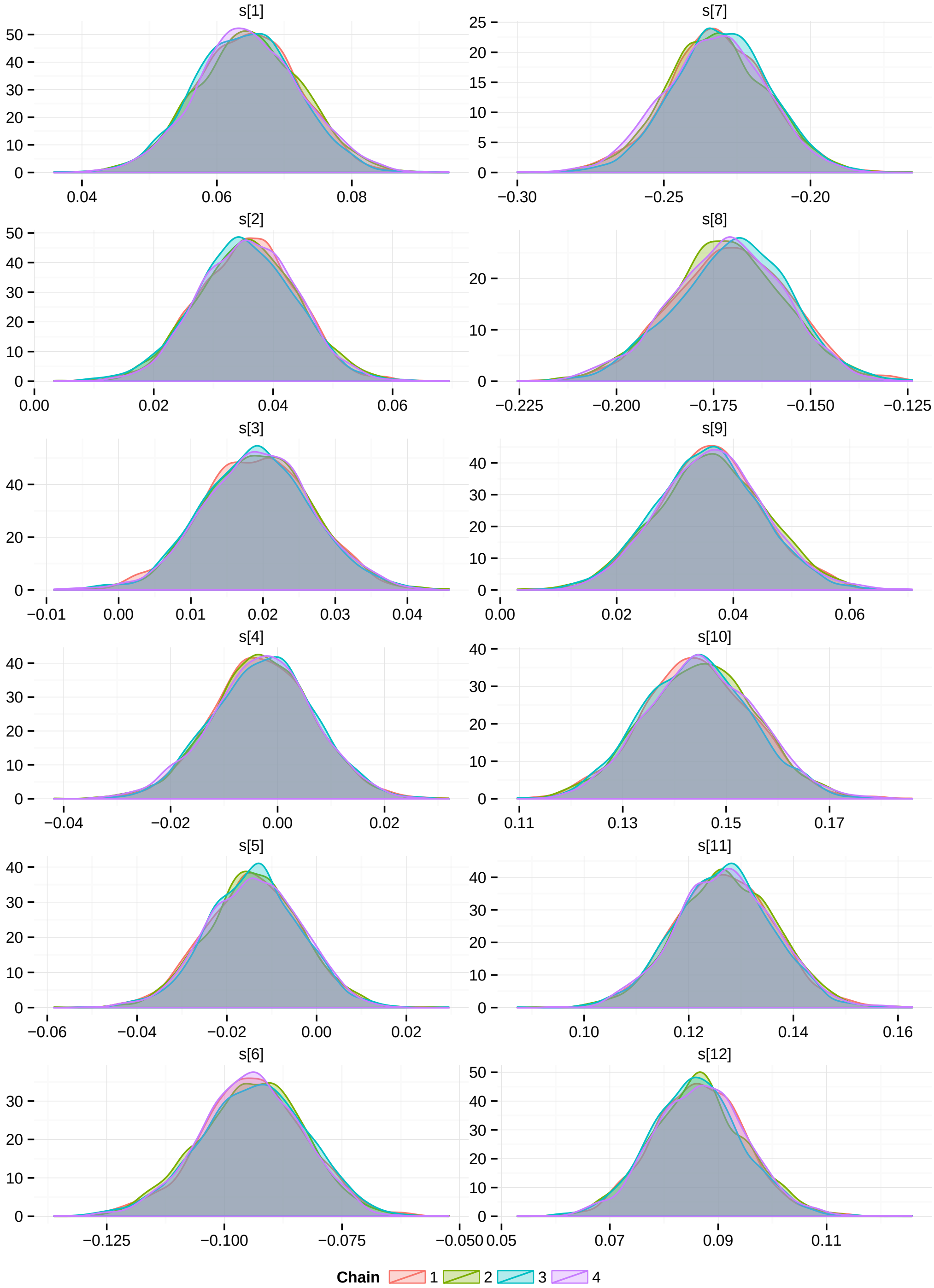

Figure 7.7: HSM/DIY Marginal Posterior Densities for Seasonal Components s_k after 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

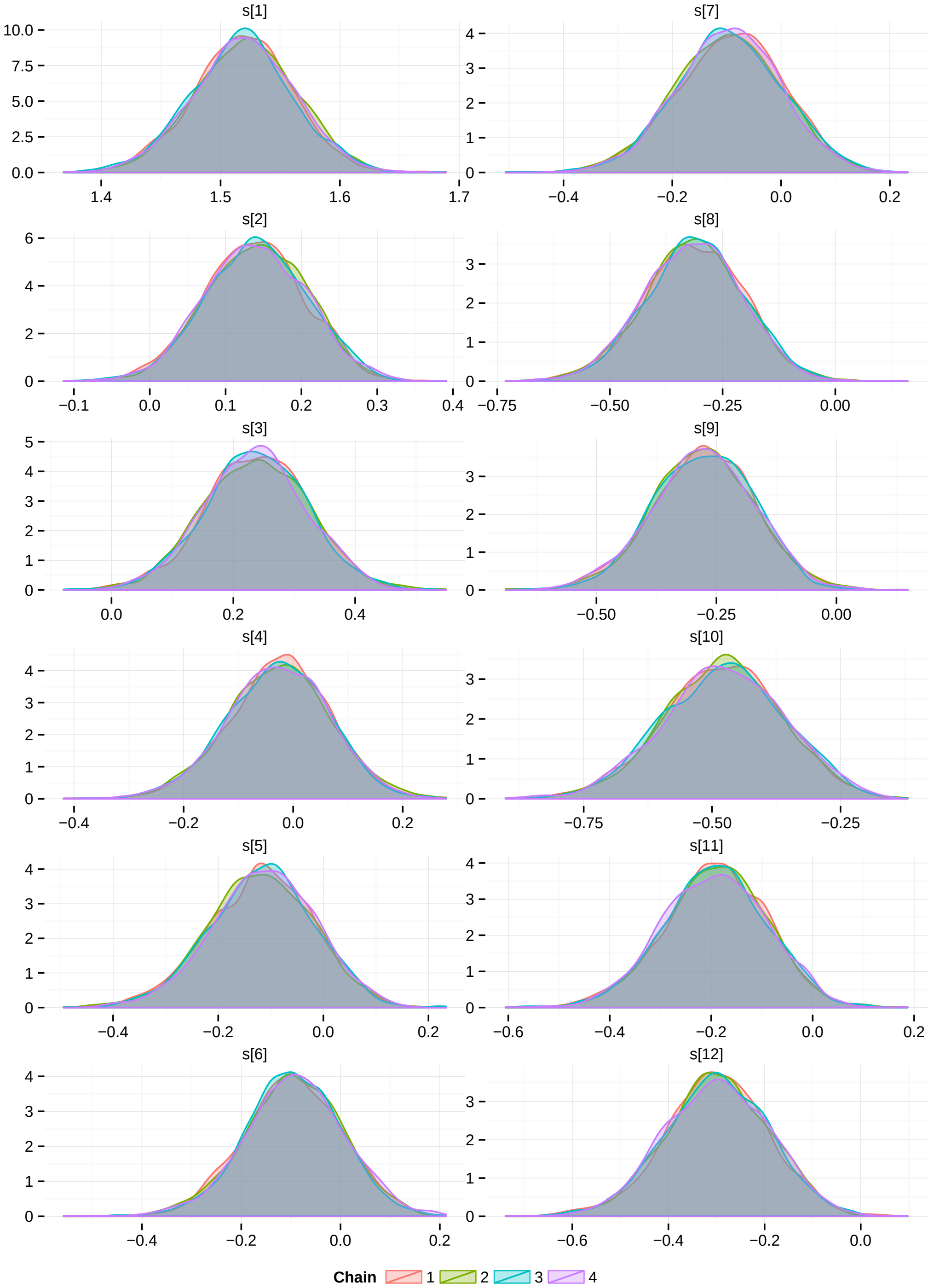

Figure 7.8: HSM/CDNOW Marginal Posterior Densities for Seasonal Components s_k after 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

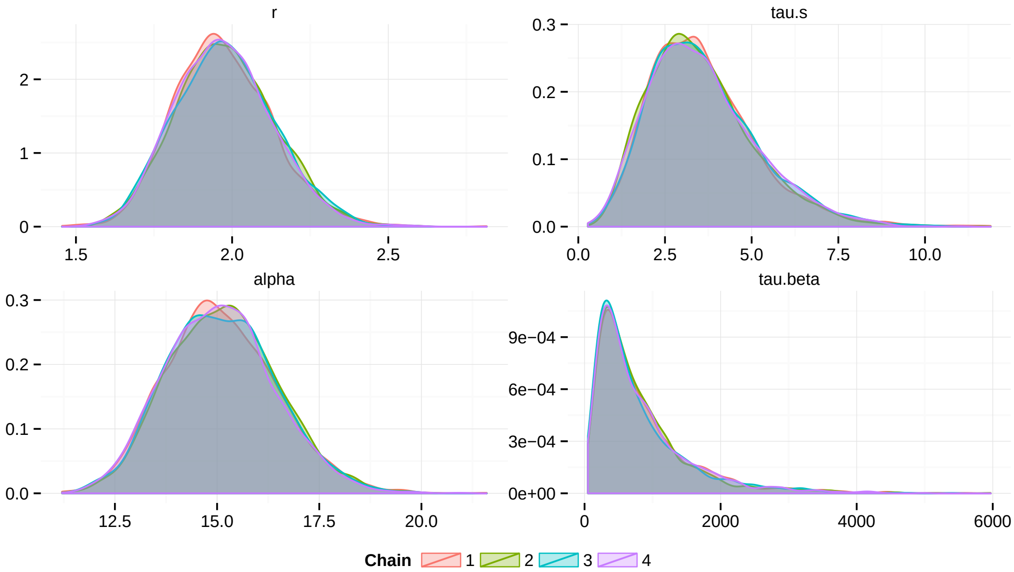

Figure 7.9: HSM/CDNOW Marginal Posterior Densities for r, α, τ_s, and τ_β after 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

7.2.2 Results and Forecast Accuracy

7.2.2.1 Seasonal Components and Individual Seasonality

The MCMC simulation for HSM yielded seasonal information on a cross-sectional level as well as on an individual customer level for all three datasets (see Tables 7.1-7.3). The estimated overall seasonal components $s_1 - s_{12}$ reflect the joint seasonal cycle. They can be interpreted as the approximate (for small values) percentage change[17] compared to the seasonal cycle average over all customers.

More important, the HSM provides individual estimates for each customer's seasonality β_i. If the seasonality measure β_i is close to 1.0 the customer follows the cross-sectional seasonal patterns $s_1 - s_{12}$. Larger $\beta_i > 1$ indicate pro-seasonal behavior with amplified purchase levels. Customers with smaller $\beta_i < 1$ exhibit attenuated seasonal behavior. A β_i near zero indicates non-seasonal behavior.

[17] This approximation is used throughout this thesis for interpreting seasonal components s_k.

Negative β_i are associated with anti-seasonal behavior. The estimates for customers' individual seasonality β_i and purchase rate λ_i are compiled in form of histograms in Figure 7.1 (DIY), Figure 7.2 (Apparel), and Figure 7.3 (CDNOW). For example, the seasonality measure β_i can be used for targeting customer groups and customer segmentation. Figure 7.10 shows an exemplary segmentation for the three retail customer bases into customers that show amplified seasonality $\beta_i \geq 1$, attenuated seasonality with β_i between 0 and 1, and anti-seasonal customers with $\beta_i < 0$.

In the following, I analyze the estimates for overall seasonal components and individual seasonality in more detail for each retailer.

Seasonal Estimates for DIY Retailer

The estimated overall seasonal components $s_1 - s_{12}$ (compare Table 7.1) for the DIY retailer indicate lower overall purchase levels from s_4 (November) through s_8 (March) and higher overall purchase levels from s_9 (April) to s_3 (October). Thus, customers' purchase frequencies are on average down approximately 23% in s_7 (February) and up approximately 14% in s_{10} (May) from their yearly mean. This overall pattern indicates that the DIY retailer's customers' activity peaks in the summer, maybe because customers use the summer holidays for home improvement or because the summer weather facilitates outdoor projects.

While overall seasonal information could have been obtained through other meth-

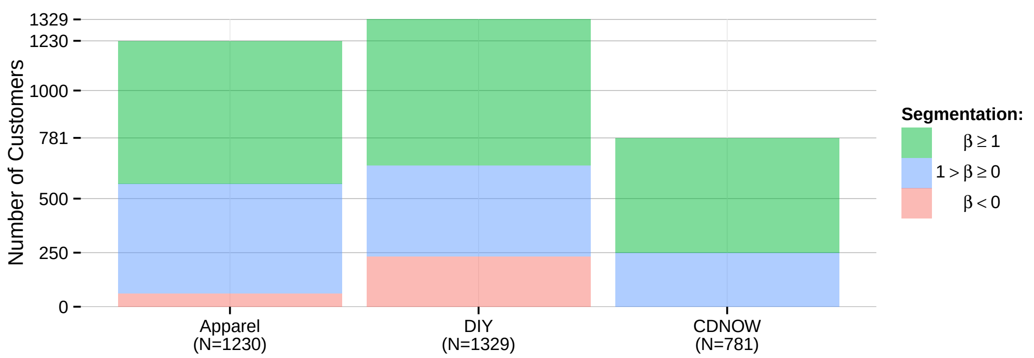

Figure 7.10: HSM/Segmentation of Customer Base by Individual Seasonality β_i for each Retailer

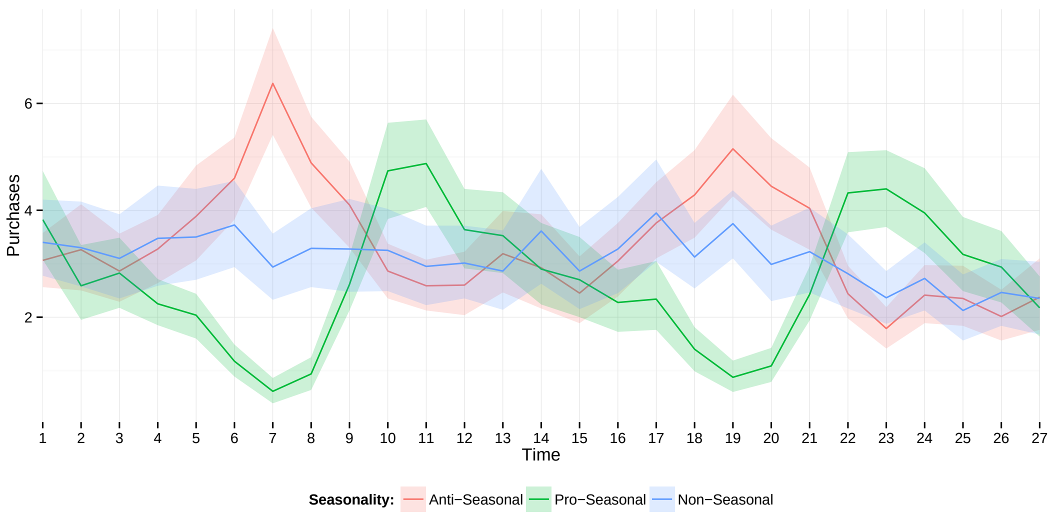

Figure 7.11: HSM/DIY Retailer Purchase History Comparison of Customer Groups Selected by Individual Seasonality Estimates β_i

ods as well, the HSM also — and more importantly — yields measures of seasonality that reflects how each customer's individual behavior relates to the behavior of other customers. The discriminative power of this measure is illustrated in Figure 7.11. From the DIY retailer's customer base I selected three groups with 50 customers each: a pro-seasonal (green), an anti-seasonal (red) and a non-seasonal (blue) customer group. The pro-seasonal group comprises the customers with the largest β_i coefficients, the anti-seasonal group consists of customers with the lowest (negative) β_i coefficients, and for the non-seasonal group, customers with β_i closest to zero were selected. The shaded area around the lines depict the 95% confidence limits of the group mean computed by a nonparametric bootstrap.

The group graphs in Figure 7.11 show clearly that the model captures individual seasonality quite well. Over time, the pro-seasonal customers' actual purchase levels are an almost exact mirror image of the anti-seasonal customers' average purchase levels, while the non-seasonal customers remain buying at fairly constant purchase levels. For example, the anti-seasonal customers' upward spikes in time periods 7 and 19 are matched by the downward spikes of the pro-seasonal customers in the same time periods. This actual behavior corresponds with the estimated overall seasonal component s_7 (February) that indicates an overall reduction in purchase levels by 23% for periods 7 and 19. Also the pro-seasonal group's upward spike in time periods 11 and 23 is consistent with the +13% increase in purchase levels implied by component s_{11} (June).

Seasonal Estimates for Apparel Retailer

The seasonal components $s_1 - s_{12}$ (compare Table 7.2) for the apparel dataset indicate, that the overall highest purchase levels are to be expected in December (+68%; $s_{12} = 0.5162$) and January (+135%; $s_1 = 0.8552$). In contrast, the lowest purchase levels are to be expected in September (-53%; $s_9 = -0.7434$) and February (-35%; $s_2 = -0.4428$). The dataset spans a 24 months time frame in total and, with a four months hold-out period, provides only eight months overlap for seasonal learning. This is reflected in the estimate for the apparel data precision parameter τ_β (1.52), which is higher than that for the DIY retailer dataset (0.44), implying that the individual β_i are more concentrated.

Figure 7.12 depicts a comparison of actual purchase histories of three customer groups selected by β_i estimates, employing the same procedure as above. While not as clear as on the DIY retailer dataset, the pro- (green) and anti-seasonal (red)

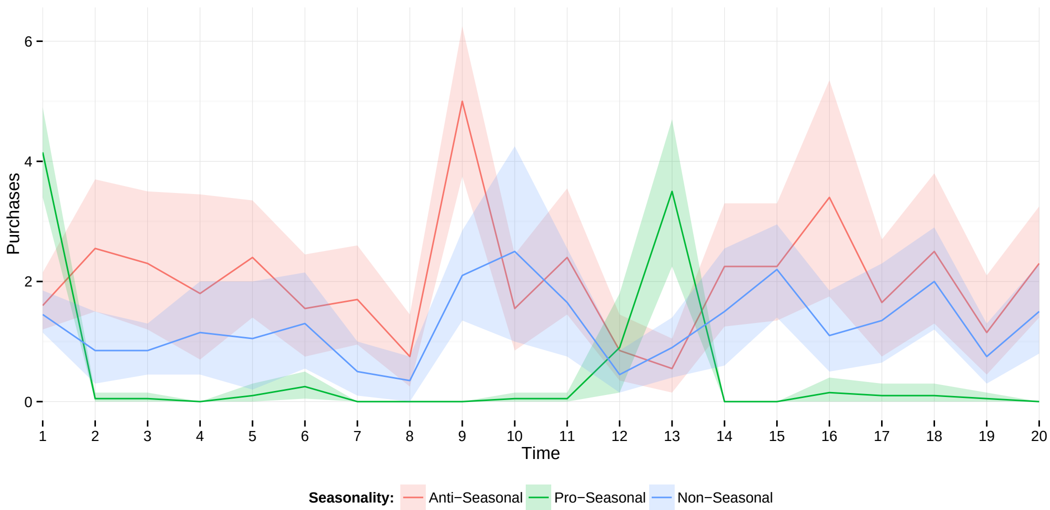

Figure 7.12: HSM/Apparel Purchase History Comparison of Customer Groups Selected by Individual Seasonality Estimates β_i

graphs mirror each other markedly over the time line. For example, the peak in pro-seasonal behavior at time periods 1 and 13 (s_1) coincides with the corresponding slump in the anti-seasonal group. Accordingly the peak in anti-seasonal activity in period 9 (s_9) is mirrored by a simultaneous slump in purchase levels of the pro-seasonal group. The non-seasonal customers (blue) show attenuated seasonal behavior that falls between the other groups.

Seasonal Estimates for CDNOW Retailer

The CDNOW dataset provides only little information for seasonal learning as it only spans 18 months in total. With four months of data reserved as hold-out that only leaves 14 months for learning and thus two month overlap for seasonal inference. Nonetheless, it is noteworthy that the model still produces individual estimates for all parameters (see Table 7.3). This a strength of hierarchical Bayesian models because the estimates in view of sparse data are pulled toward their higher level means.

Due to the lack of seasonal overlap, the individual estimates for β_i are highly concentrated ($\tau_\beta = 811$) around one (compare the histogram in Figure 7.3). This is reflected in the group comparison in Figure 7.13, where the model could only

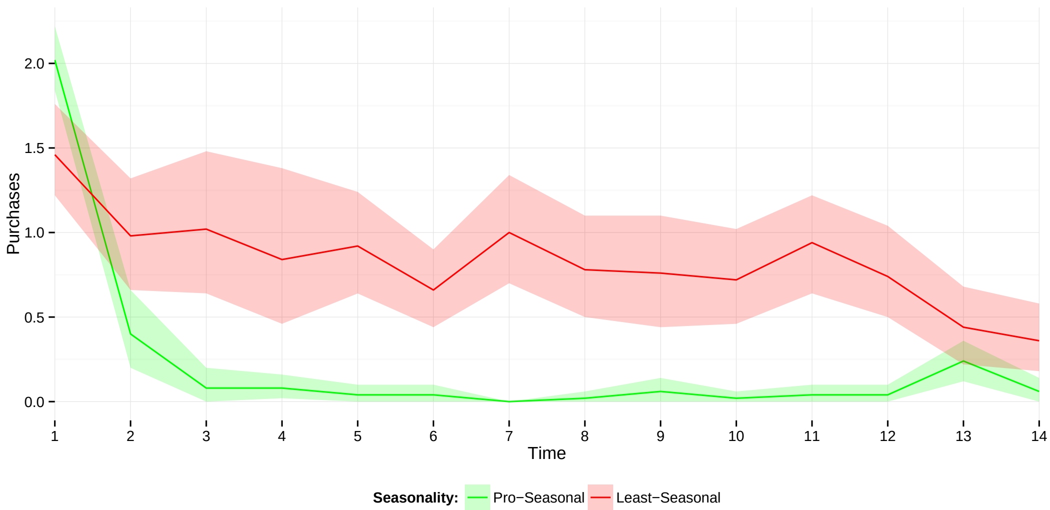

Figure 7.13: HSM/CDNOW Purchase History Comparison of Customer Groups Selected by Individual Seasonality Estimates β_i

discern pro-seasonal (green) and least-seasonal (red) customers as no values for β_i were near zero or even negative. The least-seasonal graph seems to vaguely mirror the pro-seasonal graph with slumps coinciding with pro-seasonal peaks for time periods 6 and 13. However, this is inconclusive and judging from the high concentration of β_i around one, the model could not isolate enough signal for estimating individual seasonality. Thus, the majority of the seasonal learning for the CDNOW dataset will come from the overall seasonal components $s_1 - s_{12}$. According to the estimates for the seasonal components $s_1 - s_{12}$, peak activity for CDNOW is in January ($s_1 = 1.5192$) through March ($s_3 = 0.2374$), while the lowest activity is in August ($s_8 = -0.3143$) and October ($s_{10} = -0.4787$).

7.2.2.2 Long- and Short-Term Predictions

In general, the HSM generates the most accurate short- and long-term forecasts of purchase levels. It is more accurate than the other models for the DIY and apparel dataset on both the individual level and overall error metrics[18]. The results for the CDNOW dataset are mixed with no clearly superior method among the HSM, the heuristic, and the SARIMA model. Holt-Winters method performs consistently worst and the SARIMA model produces comparable results to the heuristic and HSM only for the CDNOW data. I will detail the results in the following sections.

Long-term prediction refers to a forecast horizon covering the full length of the hold-out period (four months), while the short-term prediction entails the forecast of the number of purchases the customer is expected to make in the first unobserved time unit (one month).

The HSM is measured against Holt-Winters method, SARIMA models, and the baseline heuristic. The baseline heuristic simply assumes that:

Customers continue to buy at their mean past purchase frequency.

[18] For a description of the error metrics and error aggregation see Section 4.4.1.

The SARIMA and Holt-Winters forecast results were obtained by using the R package forecast version 5.4 (Hyndman 2014). The SARIMA and Holt-Winters models were fitted to each customer individually, estimating the model parameters with the algorithms described in Hyndman and Khandakar (2008).

Results for DIY Retailer

Tables 7.4 and 7.5 compile the long- and short-term results for the DIY retailer. The DIY retailer dataset spans the largest time frame (31 months) of the examined retailers and should therefore lend itself best to seasonal learning. The results support this. The HSM outperforms the other methods both in the short- and long-term on all error metrics. It is interesting to note that Holt-Winters method performs consistently worst while the long-term SARIMA forecasts are more accurate than the heuristic on the median aggregated metrics (*median MSE*: 2.520 vs. 2.536). Then again, the heuristic provides better long-term accuracy than the SARIMA model in terms of MAE/MAD (1.696 vs. 1.702) and *mean MSE* (5.607 vs. 5.948). Also, the heuristic generally generates less errors than the SARIMA model for short-term forecasts.

Results for Apparel Retailer

The apparel retailer's dataset spans 24 months of data which provides more overlap than the CDNOW dataset but less than the DIY retailer's. The forecast results are shown in Tables 7.6 and 7.7. Again, the HSM outperforms the other forecast methods on all metrics in both the short- and long-term, although in some cases only by a small margin. The heuristic is more accurate than the SARIMA model and Holt-Winters method for short-term forecasts, but compared to the SARIMA model the heuristic's long-term forecast results are mixed. The SARIMA model

Table 7.4: DIY Retailer - HSM Long-Term Prediction Accuracy

	Holt-Winters	**SARIMA**	**Heuristic**	**HSM**
MAE/MAD	2.335	1.702	1.696	1.636
Mean MSE	11.751	5.948	5.607	5.518
Median MSE	4.550	2.520	2.536	2.324
Mean RMSE	2.717	1.957	1.940	1.894
Median RMSE	2.133	1.587	1.592	1.524

Table 7.5: DIY Retailer - HSM Short-Term Prediction Accuracy

	Holt-Winters	SARIMA	Heuristic	HSM
MAE/MAD	10.293	5.885	1.647	1.634
Mean MSE	112.750	39.451	5.182	5.107
Median MSE	114.993	36.894	1.494	1.491
Mean RMSE	10.293	5.885	1.647	1.634
Median RMSE	10.723	6.074	1.222	1.221

is more accurate in terms of MAE/MAD (0.887 vs. 0.897) and *median MSE* (0.563 vs. 0.610), while the heuristic produces less error according to *mean MSE* (2.514 vs. 2.732). Holt-Winters method produces the least accurate forecasts in the long-term but generates better forecasts than the SARIMA model in the short-term (MAE/MAD: 1.449 vs. 1.708, *mean MSE*: 3.312 vs. 3.948 and *median MSE*: 1.855 vs. 3.063).

Results for CDNOW

The CDNOW dataset only spans 18 months in total. With four months of data reserved as hold-out, only 14 months are available for parameter learning, implying only a two months overlap for seasonal inference. Holt-Winters method did not yield any results as the overlap was too small — this is denoted as *NA*. The forecast results for the CDNOW dataset are compiled in Tables 7.8 and 7.9. The short and long-term results show no clear pattern. The SARIMA model, the heuristic and the HSM are very close without any model setting itself apart. For example, in the long-term forecast in regard to *mean MSE*, the HSM (0.089) is more accurate than the SAMIRA model (0.122) and the heuristic (0.125). But in terms of MAE/MAD the SARIMA model (0.131) is more accurate than the HSM (0.170) and the heuristic (0.206). In the short-term the SARIMA model is inferior to both the heuristic and HSM on all metrics. Then again, in terms of *mean MSE*, the HSM (0.146) is more accurate than the heuristic (0.157), while in regard to MAE/MAD this reverses with the heuristic (0.222) generating slightly more accurate predictions than the HSM (0.223).

Table 7.6: Apparel - HSM Long-Term Prediction Accuracy

	Holt-Winters	SARIMA	Heuristic	HSM
MAE/MAD	1.254	0.887	0.897	0.884
Mean MSE	4.488	2.732	2.514	2.509
Median MSE	1.374	0.563	0.610	0.536
Mean RMSE	1.570	1.104	1.105	1.104
Median RMSE	1.172	0.750	0.781	0.732

Table 7.7: Apparel - HSM Short-Term Prediction Accuracy

	Holt-Winters	SARIMA	Heuristic	HSM
MAE/MAD	1.449	1.708	0.874	0.786
Mean MSE	3.312	3.948	2.367	2.306
Median MSE	1.855	3.063	0.250	0.072
Mean RMSE	1.449	1.708	0.874	0.786
Median RMSE	1.362	1.750	0.500	0.269

Table 7.8: CDNOW - HSM Long-Term Prediction Accuracy

	Holt-Winters	SARIMA	Heuristic	HSM
MAE/MAD	*NA*	0.131	0.206	0.170
Mean MSE	*NA*	0.122	0.125	0.089
Median MSE	*NA*	0.000	0.005	0.008
Mean RMSE	*NA*	0.157	0.225	0.194
Median RMSE	*NA*	0.000	0.071	0.090

Table 7.9: CDNOW - HSM Short-Term Prediction Accuracy

	Holt-Winters	SARIMA	Heuristic	HSM
MAE/MAD	*NA*	0.347	0.222	0.223
Mean MSE	*NA*	0.197	0.157	0.146
Median MSE	*NA*	0.082	0.005	0.013
Mean RMSE	*NA*	0.347	0.222	0.223
Median RMSE	*NA*	0.286	0.071	0.113

7.2.2.3 Prediction of Future Best Customers

The HSM provides equal or better predictions of the future top 10% and 20% of customers for the DIY and apparel retail data compared to the heuristic, SARIMA and Holt-Winters method. The predictions by the HSM are also equal or better than the heuristic for the customer base of CDNOW. Only in forecasting the top 20% of customers does the SARIMA model slightly outperform the HSM. Holt-Winters method consistently produces the largest errors. Even though the HSM performs best, the differences are very slight between the HSM, SARIMA, and the heuristic. Below, I will detail the results.

The ability to identify future high- and low-value customers is tested by rank ordering customers according to the HSM's forecast of purchase frequencies for the hold-out period. The actual future top 10% (20%) of customers (positives) are determined by their observed purchase frequency in the hold-out period. The prediction of an individual customer can be correct (true) or incorrect (false).

Rank ordering provides a different view on forecast accuracy compared to the error metrics in the previous section. For example, even if forecast errors are quite large for a particular method, customer purchase levels relative to other customers could be captured accurately. The methods are benchmarked against the following baseline heuristic:

> *The past 10% (20%) best customers will also be the future 10% (20%) best customers.*

Results for DIY Retailer

The prediction accuracy for the top 10% and 20% of the DIY retailer's customer base are presented in Table 7.10. The accuracy of all methods in determining the future top 20% customers is very close, ranging from 84.95% (HSM), 84.95% (SARIMA), 84.80% (heuristic) to 82.09% (Holt-Winters). A similar picture results for the top 10% with accuracy ranging from 90.97% (HSM), 90.97% (heuristic), 90.67% (SARIMA) to 89.16% (Holt-Winters).

Table 7.10: DIY Retailer - Accuracy of Top 10% (20%) Customer Prediction

(%)	Holt-Winters		SARIMA		Heuristic		HSM	
True Positives:	45.46	(55.09)	53.03	(62.26)	54.55	(61.89)	54.55	(62.26)
True Negatives:	93.98	(88.82)	94.82	(90.60)	94.99	(90.51)	94.99	(90.60)
False Negatives:	54.54	(44.91)	46.97	(37.74)	45.45	(38.11)	45.45	(37.74)
False Positives:	6.02	(11.18)	5.18	(9.40)	5.01	(9.49)	5.01	(09.40)
Error:	10.84	(17.91)	9.33	(15.05)	9.03	(15.20)	9.03	(15.05)
Accuracy:	89.16	(82.09)	90.67	(84.95)	90.97	(84.80)	90.97	(84.95)

Table 7.11: Apparel - Accuracy of Top 10% (20%) Customer Prediction

(%)	Holt-Winters		SARIMA		Heuristic		HSM	
True Positives:	40.65	(48.78)	52.85	(59.35)	52.85	(59.76)	53.66	(60.16)
True Negatives:	93.41	(87.19)	94.76	(89.84)	94.76	(89.94)	94.85	(90.04)
False Negatives:	59.35	(51.22)	47.15	(40.65)	47.15	(40.24)	46.34	(39.84)
False Positives:	6.59	(12.81)	5.24	(10.16)	5.24	(10.06)	5.15	(09.96)
Error:	11.87	(20.49)	9.43	(16.26)	9.43	(16.10)	9.27	(15.93)
Accuracy:	88.13	(79.51)	90.57	(83.84)	90.57	(83.90)	90.73	(84.07)

Table 7.12: CDNOW - Accuracy of Top 10% (20%) Customer Prediction

(%)	Holt-Winters		SARIMA		Heuristic		HSM	
True Positives:	*NA*		39.74	(44.87)	41.03	(42.31)	41.03	(42.95)
True Negatives:	*NA*		93.31	(86.24)	93.46	(85.60)	93.46	(85.76)
False Negatives:	*NA*		60.26	(55.13)	58.97	(57.69)	58.97	(57.05)
False Positives:	*NA*		6.69	(13.76)	6.54	(14.40)	6.54	(14.24)
Error:	*NA*		12.04	(22.02)	11.78	(23.05)	11.78	(22.79)
Accuracy:	*NA*		87.96	(77.98)	88.22	(76.95)	88.22	(77.20)

Results for Apparel Retailer

The prediction accuracy for the apparel customer base is presented in Table 7.11. The methods' accuracy in determining the future top 20% customers is very close, ranging from 84.07% (HSM), 83.90% (heuristic), 83.84% (SARIMA) to 79.51% (Holt-Winters). A similar picture results for the top 10% with accuracy ranging from 90.73% (HSM), 90.57% (heuristic), 90.57% (SARIMA) to 88.13% (Holt-Winters).

Results for CDNOW

Table 7.12 compiles the rank-order results for the CDNOW dataset. The accuracy of the HSM is equal (Top 10%: 88.22%) or better (top 20%: 77.20% vs. 76.95%) than the accuracy of the heuristic. Likewise, the HSM is more accurate in predicting the top 10% compared to the SARIMA model (88.22% vs. 87.96%). Only in identifying the top 20% customer, does the SARIMA model slightly outperform the seasonal model (77.98% vs. 77.20%).

7.3 Discussion of Results

The HSM operates under the "always a share" assumption. This model contributes to the literature stream of "probability models for customer base analysis" (Fader and Hardie 2009) and is unique in that it combines (1) individual purchase rates and cross-sectional heterogeneity, (2) multiplicative seasonal effects, (3) individual customers' seasonal behavior and group-level seasonal behavior, and (4) a measure of individual seasonality.

The model is based on a discretely sampled inhomogeneous Poisson counting process that features a multiplicative time-varying seasonal co-variate submodel. Gibbs sampling in a MCMC framework was used to obtain parameter estimates and forecasts for the hold-out period. In this respect the model extends the NBD family and related variants (e.g., Schmittlein, Morrison, and Colombo 1987; Fader, Hardie, and Lee 2005a; Jerath, Fader, and Hardie 2011; Bemmaor and Glady 2012; Mzoughia and Limam 2014; Glady, Lemmens, and Croux 2015) by including individual and cross-sectional seasonal effects. Also, it generalizes post-hoc seasonal adjustment (Zitzlsperger, Robbert, and Roth 2009) and dummy-variable approaches (Schweidel and Knox 2013) to improve forecast accuracy in this literature stream.

The model yields a measure of individual seasonality that indicates how strongly the customer follows the cross-sectional seasonality, if he purchases anti-cyclically, or if the customer's behavior is non-seasonal. The analysis of the HSM shows that it captures both individual and cross-sectional seasonality quite well (see Section 7.2.2.1). The results of targeting customer groups and segmenting the customer base according to the model's seasonal estimates demonstrate that this measure has a high discriminative power, even with datasets that feature only moderate seasonal overlap (as shown for example in Figure 7.11).

The targeting of pro-seasonal, non-seasonal and anti-seasonal customers is particularly useful to a marketing manager to improve the timing and effectiveness of marketing actions. For example, cross-selling initiatives should be timed to coincide with peaks in customer activity. As the time periods with peaks differ substantially between pro- and anti-seasonal customers groups, the targeting by seasonality could raise success rate of such campaigns. Executives could use this information for customer portfolio management, mitigating risk by evening out seasonal peaks and seasonal slumps by attracting or rewarding customers, who fit the desired seasonal profile.

The empirical validation provided further insights into the forecast accuracy of the seasonal model compared to a simple heuristic, SARIMA models, and Holt-Winters method. The analysis of the long- and short-term forecasts shows that the HSM consistently outperforms the other methods on the DIY and apparel data. This holds true for both the overall and aggregated individual level error metrics. The forecast accuracy for CDNOW's customers does not provide a clear picture. Neither the HSM, the SARIMA method, nor the heuristic sets itself apart from the other forecast methods. This is in part expected as the CDNOW data contains the least seasonal overlap among the retailers, allowing only reduced seasonal learning. Holt-Winters method consistently produced the highest errors.

The HSM's ability to identify future high-value customers follows the above pattern. For both the DIY and apparel data, the forecasts are equal or better than the heuristic, SARIMA, and Holt-Winters method. The results for CDNOW's customers are again mixed. While the HSM is equal or better than the heuristic, the SARIMA model is slightly more accurate in identifying the future top 20% of the customer base.

In general, the HSM improves on the heuristic, SARIMA and Holt-Winters methods in the managerial tasks simulated in this benchmark. The other methods only provide competitive forecast accuracy for CDNOW's customer. This might be in part due to the minimal seasonal overlap of CDNOW's data, but also because it contains a relatively high number of customers who did not purchase recently and might have dropped out. The ability to improve forecasts for customer bases that exhibit substantial customer drop-out is the motivation for developing a hierarchical Bayesian seasonal model with drop-out (HSMDO) in the next chapter.

Chapter 8

Hierarchical Bayesian Seasonal Effects Model with Drop-Out

The hierarchical Bayesian seasonal effects model with drop-out (HSMDO) I develop in this chapter includes a hierarchical customer lifetime model under the "buy 'til you die" assumption. This replaces the "always a share" assumption underlying the HSM.

The HSMDO aims at (1) improving forecast accuracy by incorporating an explicit customer lifetime model, (2) relating individual and cross-sectional drop-out rates, (3) relating individual purchase rates to cross-sectional heterogeneity, (4) including multiplicative seasonal effects, (5) relating individual customers' seasonal behavior to group-level seasonal behavior, (5) yielding a measure of individual seasonality, and (6) providing a probabilistic measure that indicates whether a customer is still active or will remain inactive for a certain period of time.

The HSMDO yields an individual measure of a customer's seasonality that accounts for customer drop-out. It can be used by practitioners to target customers, segment the customer base and improve the efficiency of marketing campaigns.

8.1 General Model Framework

The model is based on a discretely sampled inhomogeneous Poisson counting process where death opportunities coincide with the granularity of the purchase frequency sampling and incorporates the following features:

1. a hierarchical purchase process that relates the distribution of individual level purchase frequencies over time to cross-sectional heterogeneity;
2. a hierarchical seasonal structure that relates individual level seasonality to cross-sectional seasonal components in a multiplicative submodel;
3. an individual estimate for each customer's seasonality that indicates how strongly he follows the cross-sectional seasonality, whether he purchases anti-seasonal or non-seasonal;
4. a hierarchical customer lifetime model that relates the distribution of individual level customer lifetimes to cross-sectional heterogeneity;
5. an individual estimate for each customer's probability of being active after the observation period: $P(alive)$;
6. a new alternative to $P(alive)$ that yields a theoretically sound and managerial relevant measure of customer activity/inactivity for finite time horizons, $P(Zero_F)$; and
7. model parameter estimation and prediction through Hybrid/Hamilton Monte Carlo Methods (Brooks et al. 2011, pp. 113-160) in form of the no-U-turn sampler (Homan and Gelman 2014).

The drop-out process I propose is closely related to previous work on the Pareto/NBD model and related variants (e.g., Schmittlein, Morrison, and Colombo 1987; Fader, Hardie, and Lee 2005a; Jerath, Fader, and Hardie 2011; Bemmaor and Glady 2012; Mzoughia and Limam 2014; Glady, Lemmens, and Croux 2015) as discussed in Section 3.1.2.

Similar to the PDO/NBD model (Jerath, Fader, and Hardie 2011), I assume discrete periodic death opportunities tied to calendar time. My model differs from the PDO/NBD in that it is based on a discretely sampled inhomogeneous Poisson counting process so that death opportunities coincide with the granularity of the purchase frequency sampling. This is at no loss of generality as the data can be re-sampled at the desired time scale before applying the model.

Ever since Schmittlein, Morrison, and Colombo (1987) introduced the $P(alive)$ metric, it has been customary for researchers to derive $P(alive)$ for basically all models that contain a form of drop-out process (Fader and Hardie 2009). $P(alive)$ indicates the likelihood that a customer has not dropped out and is still active after the observation period. Unfortunately, $P(alive)$ pertains to an infinite time horizon, posing problems both for empirical validation and managerial relevance (Wübben and von Wangenheim 2008; Fader, Hardie, and Shang 2010). I derive an alternative metric $P(Zero_F)$ for the HSMDO that solves these issues and yields a theoretically sound probabilistic measure for customer inactivity in finite time horizons. The conceptual idea, theoretical background, managerial implications, and mathematical derivation is presented in Section 8.3, p. 136.

For model parameter estimation I use Hybrid/Hamilton Monte Carlo Sampling (Brooks et al. 2011, pp. 113-160) in form of the no-U-turn sampler (Homan and Gelman 2014) instead of Gibbs sampling that was used for the DMPT and HSM. Hamilton Monte Carlo sampling substantially increases the effective sample size obtained in the same amount of computational time (Carpenter et al. 2015).

Before deriving the joint model likelihood, the conditional expected purchase levels, as well as the $P(alive)$ and $P(Zero_F)$ metrics, I present the model assumptions and the mathematical notation in the following order:

1. the customer lifetime model with periodic drop-out opportunities: shifted-geometric beta mixture;
2. during lifetime: individual customers purchase according to a Poisson time-varying covariate gamma mixture;
3. the multiplicative seasonal effects co-variate model; and
4. prior distributions, hyperparameters and parameter space.

8.1.1 Customer Lifetime Model with Periodic Drop-out Opportunities: Shifted-Geometric Beta Mixture

The lifetime model assumes that customer i's, with $i \in \{1..N\}$, relationship with the firm can be characterized as first being alive for some period of time τ_i, then dropping out and becoming permanently inactive ("dead"). Each customers unobserved lifetime of τ_i follows a shifted-geometric distribution with drop-out probability per time unit of p_i:

$$\tau_i \sim \quad SGeometric(p_i) \quad \overset{\text{p.m.f.}}{\rightleftharpoons} p_i(1-p_i)^{\tau_i - 1} \tag{8.1}$$

$$p_i \sim \quad Beta(a,b) \quad \overset{\text{p.d.f.}}{\rightleftharpoons} \frac{p_i^{a-1}(1-p_i)^{b-1}}{B(a,b)} \tag{8.2}$$

$$B(a,b) = \quad \int_0^1 y^{a-1}(1-y)^{b-1}\,\mathrm{d}y \quad = \Gamma(a)\Gamma(b)/\Gamma(a+b) \tag{8.3}$$

Cross sectional heterogeneity in drop-out probabilities p_i follows a Beta distribution with shape parameters $a,b>0$. $B(a,b)$ is the beta function or Euler integral of the first kind. In the context of the discretely sampled Poisson counting process, starting in time period one, a drop-out event at time τ_i means that customer i drops-out immediately after the τ_i's time period and remains inactive in all future time periods starting with $\tau_i + 1$.

Let x_{ij} denote the observed number of purchases customer i makes in time period j, then a customer drop-out at τ_i implies that $x_{i,\tau_i+k} = 0$ for all $k > 0$. All customers are assumed to be alive in time period one.

8.1.2 Poisson Covariate Gamma Mixture

While alive each customer i with $i \in \{1..N\}$ is assumed to buy at an individual rate $\lambda_i e^{\beta_i s_{k(j)}}$ in time period j governed by a Poisson purchase process[19]. The individid-

[19] This process can be seen as an discrete sampled inhomogeneous Poisson counting process with $N(j) - N(j-1) = \Delta N_j \sim Poisson(\int_{j-1}^{j} \lambda(t)dt)$, where $\lambda(t)$ is constant between $(j-1, j]$ and therefore $\Delta N_j \sim Poisson(\lambda(j))$.

ual purchase histories are organized so that an individual customer i's purchases in time period j are denoted $x_{ij} \in \mathbb{Z}_{\geq 0}$.

The heterogeneity across customers' base purchase rates λ_i is captured by a gamma distribution. The parameters of the group level gamma distribution are the scale parameter $r > 0$ and the shape parameter $\alpha > 0$:

$$x_{ij} \sim \quad Poisson(\lambda_i e^{\beta_i s_{k(j)}}) \overset{\text{p.m.f.}}{\rightleftharpoons} e^{-\lambda_i e^{\beta_i s_{k(j)}}} \frac{(\lambda_i e^{\beta_i s_{k(j)}})^{x_{ij}}}{\Gamma(x_{ij}+1)} \tag{8.4}$$

$$\lambda_i \sim \quad Gamma(r, \alpha) \overset{\text{p.d.f.}}{\rightleftharpoons} e^{-\lambda_i \alpha} \lambda_i^{r-1} \frac{\alpha^r}{\Gamma(r)} \tag{8.5}$$

The time-varying random-effects submodel $\beta_i s_{k(j)}$ describes the seasonal structure of the model. The term depends both on time j and customer i and reflects both the joint seasonal structure on a group level as well as individual level seasonal effects as explained below.

8.1.3 Hierarchical Seasonal Effects Submodel

While alive, the individual level seasonality of each customer i is captured by a multiplicative random effect β_i. It determines the influence of the overall seasonal components $s_{k(j)}$ on each customer's purchase rate λ_i. This results in Poisson purchase rate $\lambda_i e^{\beta_i s_{k(j)}}$ for customer i in time period j. Cross sectional heterogeneity in β_i is assumed to be distributed normal with mean one and precision τ_β.

Let s_k denote the k^{th} seasonal component with $k \in \{1..K\}$. Each point in time j is linked[20] to one of the K seasonal components by $k(j) = ((j-1) \bmod K) + 1$. The (uncentered) seasonal components $\hat{s}_k$ are assumed to be exchangeable and distributed normal with mean zero and precision τ_s. The components $\hat{s}_k$ are centered at zero, so that s_k reflect the (approximate) seasonal percentage change compared to average purchases levels.

[20] The model allows for any form of time-varying multiplicative co-variates and is derived in full generality. For the general case the link function may be set to the identity function $k(j) = j$

I constructed the model to comprehensively capture seasonal behavior. For example, if an individual customer i exhibits amplified (pro)seasonal behavior his seasonality coefficient would be $\beta_i > 1$. If a customer shows an attenuated seasonal pattern, compared to the cross-sectional seasonal components, then his coefficient would be $0 < \beta_i < 1$. If a customer purchases fairly constantly at his individual purchase rate λ_i, implying non-seasonal behavior, his coefficient β_i would be near or at zero. Finally, if a customer shows anti-seasonal purchase patterns, behaving opposite to the overall seasonal behavior, his coefficient would be $\beta_i < 0$.

The structure of the seasonal submodel can then be decomposed and summarized as follows:

$$\hat{s}_k \sim \quad Normal(0, \tau_s) \quad \overset{\text{p.d.f.}}{\rightleftharpoons} \sqrt{\frac{\tau_s}{2\pi}}\, e^{-\tau_s \hat{s}_k^2/2} \tag{8.6}$$

$$s_k = \quad \hat{s}_k - \tfrac{1}{K}\sum \hat{s}_k \tag{8.7}$$

$$\beta_i \sim \quad Normal(1, \tau_\beta) \quad \overset{\text{p.d.f.}}{\rightleftharpoons} \sqrt{\frac{\tau_\beta}{2\pi}}\, e^{-\tau_\beta (\beta_i - 1)^2/2} \tag{8.8}$$

$$k(j) = \quad ((j-1) \bmod K) + 1 \tag{8.9}$$

In a Bayesian hierarchical model one can include all K seasonal components s_k and N seasonal parameters β_i as they are drawn from distributions with finite precisions and that group level information allows for identification. The cross-sectional parameters of the seasonal model are $\tau_s > 0$ and $\tau_\beta > 0$.

8.1.4 Priors, Hyperparameters and Parameter Space

The precision parameters (τ_s, τ_β) and the parameters of the Poisson-gamma mixture (r, α) are given vague non-informative gamma priors through the choice of hyperparameters. The parameters a and b for the beta distributions are reparameterized in terms of a mean parameter ϕ_μ with a uniform (between 0 and 1) prior and a total count parameter ϕ_c with a weakly informative Pareto prior (compare Gelman et al. 2003, p. 128). Together with the corresponding parameter transformations these prior distributions complete the model framework:

$$\tau_s, \tau_\beta, r, \alpha \sim \quad Gamma(\varepsilon, \varepsilon) \tag{8.10}$$

$$\phi_\mu \sim \quad Uniform(0,1) \tag{8.11}$$

$$\phi_c \sim \quad Pareto(\varepsilon_{min}, \varepsilon_\alpha) \tag{8.12}$$

$$a = \phi_\mu \phi_c \quad b = \phi_c(1-\phi_\mu) \tag{8.13}$$

The values used in this thesis for the hyperparameters ε, ε_{min}, ε_α are 0.001, 0.1 and 1.5, respectively. Let N be the number of customers and K the number of seasonal components then the model contains $3N+K+6$ parameters of interest. The term $3N+K+6$ results from N individual purchase frequencies parameters λ_i, N individual seasonality measures β_i, N individual drop-out probabilities p_i, K seasonal components s_k, and the six cross-sectional parameters r, α, τ_s, τ_β, ϕ_c and ϕ_μ, while a and b are just transformations of ϕ_μ and ϕ_c.

8.2 Mathematical Derivation of the HSMDO

The individual purchase histories are organized so that an individual customer i's purchases in time period j are denoted x_{ij}. The notation used to present the observed information about a customer i is $(X_{ij} = x_{ij},\ t_i,\ T)$, where x_{ij} is the number of transactions observed by customer i in time period j with $1 \leq j \leq T$, and T is the number of observed time periods. The last time period when the observed number of transactions was larger than zero is denoted t_i. This implies that $x_{it_i} > 0$ and $x_{im} = 0$ for all $t_i < m \leq T$. This does not imply that the customer dropped out at t_i. Let τ_i be the unobserved time period customer i drops out[21]. Then the customer might have already dropped out at any of the time periods $T \geq \tau_i \geq t_i$ or he is still alive and drops out in the future with $\tau_i > T$.

[21] A drop-out event at time τ_i means that customer i drops-out immediately after the τ_i's time period and remains inactive in all future time periods starting with $\tau_i + 1$. This implies that $x_{i\tau_i+k} = 0$ for all $k > 0$. All customers are assumed to be alive in time period one.

8.2.1 Derivation of Model Likelihood

First, I derive the individual level likelihood function. For better readability I will drop the customer indices i, set $s_{k(j)} = s_j$ and let s denote the set of s_k with $k \in \{1..K\}$. Lets first suppose the time period $\tau \geq t$, when the customer drops out, is known. This implies no further purchases from $\tau+1$ onwards[22]. Then conditional on $(\tau,\ X_j = x_j,\ t,\ T,\ s)$, and using Equation (8.4) the individual level likelihood is:

$$L(\lambda,\beta|\tau,X_j = x_j,t,T,s) = \prod_{j=1}^{t} Poisson(x_j|\lambda e^{\beta s_j}) \prod_{j=t+1}^{min(\tau,T)} Poisson(0|\lambda e^{\beta s_j}) \tag{8.14}$$

Taking the expectation over τ, using Equation (8.1), partitioning the sum and using that x_j=0 for $j > t$ yields:

$$\begin{aligned} &L(\lambda,\beta,p|X_j = x_j,t,T,s) \\ &= \sum_{\tau=t}^{\infty}\Big[SGeometric(\tau|p)\prod_{j=1}^{t} Poisson(x_j|\lambda e^{\beta s_j}) \prod_{j=t+1}^{min(\tau,T)} Poisson(0|\lambda e^{\beta s_j})\Big] \\ &\qquad = \sum_{\tau=T+1}^{\infty}\Big[SGeometric(\tau|p)\prod_{j=1}^{T} Poisson(x_j|\lambda e^{\beta s_j})\Big] + \\ &\qquad\qquad \sum_{\tau=t}^{T}\Big[SGeometric(\tau|p)\prod_{j=1}^{\tau} Poisson(x_j|\lambda e^{\beta s_j})\Big] \end{aligned} \tag{8.15}$$

Using $\sum_{\tau=T+1}^{\infty} SGeometric(\tau|p) = (1-p)^T$ to eliminate the infinite sum, substituting and eliminating common factors that only depend on observed data, the

22 The customer is known to have been active in time period t, as he made at least one purchase in t and thus $P(\tau < t|t) = 0$.

likelihood (8.15) is further simplified[23]:

$$
\begin{aligned}
&L(\lambda,\beta,p|X_j=x_j,t,T,s)\\
&\quad=(1-p)^T\prod_{j=1}^{T}Poisson(x_j|\lambda e^{\beta s_j})+\sum_{\tau=t}^{T}\left[(1-p)^{\tau-1}p\prod_{j=1}^{\tau}Poisson(x_j|\lambda e^{\beta s_j})\right]\\
&\quad=(1-p)^T\prod_{j=1}^{T}\frac{e^{-\lambda e^{\beta s_j}}\left(\lambda e^{\beta s_j}\right)^{x_j}}{\Gamma\left(x_j+1\right)}+\sum_{\tau=t}^{T}\left[(1-p)^{\tau-1}p\prod_{j=1}^{\tau}\frac{e^{-\lambda e^{\beta s_j}}\left(\lambda e^{\beta s_j}\right)^{x_j}}{\Gamma\left(x_j+1\right)}\right]\\
&\quad=(1-p)^T\prod_{j=1}^{T}e^{-\lambda e^{\beta s_j}}\left(\lambda e^{\beta s_j}\right)^{x_j}+\sum_{\tau=t}^{T}\left[(1-p)^{\tau-1}p\prod_{j=1}^{\tau}e^{-\lambda e^{\beta s_j}}\left(\lambda e^{\beta s_j}\right)^{x_j}\right]\\
&\quad=(1-p)^T\lambda^{\sum_{j=1}^{t}x_j}e^{-\lambda\sum_{j=1}^{T}e^{\beta s_j}}e^{\beta\left(\sum_{j=1}^{t}s_jx_j\right)}+\\
&\qquad\qquad\sum_{\tau=t}^{T}\left[p(1-p)^{\tau-1}\lambda^{\sum_{j=1}^{t}x_j}e^{-\lambda\sum_{j=1}^{\tau}e^{\beta s_j}}e^{\beta\left(\sum_{j=1}^{t}s_jx_j\right)}\right]\\
&\quad=\lambda^{\sum_{j=1}^{t}x_j}e^{\beta\left(\sum_{j=1}^{t}s_jx_j\right)}\left[(1-p)^Te^{-\lambda\sum_{j=1}^{T}e^{\beta s_j}}+\sum_{\tau=t}^{T}p(1-p)^{\tau-1}e^{-\lambda\sum_{j=1}^{\tau}e^{\beta s_j}}\right]
\end{aligned}
\tag{8.16}
$$

Sampling directly from a sample-likelihood that is based on Equations (8.16), (8.2) and (8.5)-(8.13) is still not very efficient. Fortunately, the parameter p can be marginalized out. This allows for better exploration of the tails of the distribution and more efficient sampling because the expectation of p does not need to be estimated through sampling. I use Equations (8.2) and (8.3) to integrate out p,

[23] For example for all τ with $t\le\tau\le T$, the products $\prod_{j=1}^{\tau}\Gamma(x_j+1)=\prod_{j=1}^{t}\Gamma(x_j+1)$ are equal, because $x_j=0$ for all $j>t$. The same logic applies to $\sum_{j=1}^{\tau}s_jx_j$ and $\sum_{j=1}^{\tau}x_j$, in both cases the sum can be trimmed to t instead of τ.

yielding:

$$
\begin{aligned}
&L(\lambda,\beta,a,b|X_j=x_j,t,T,s)\\
&=\int_0^1 Beta(p|a,b)\lambda^{\sum_{j=1}^t x_j} e^{\beta\left(\sum_{j=1}^t s_j x_j\right)}\\
&\qquad\qquad \left[(1-p)^T e^{-\lambda\sum_{j=1}^T e^{\beta s_j}}+\sum_{\tau=t}^T p(1-p)^{\tau-1} e^{-\lambda\sum_{j=1}^{\tau} e^{\beta s_j}}\right]\mathrm{d}p\\
&=\frac{\lambda^{\sum_{j=1}^t x_j} e^{\beta\left(\sum_{j=1}^t s_j x_j\right)}}{B(a,b)}\\
&\qquad \left[B(a,b+T)e^{-\lambda\sum_{j=1}^T e^{\beta s_j}}+\sum_{\tau=t}^T B(1+a,b+\tau-1)e^{-\lambda\sum_{j=1}^{\tau} e^{\beta s_j}}\right]
\end{aligned}
\tag{8.17}
$$

Restoring the indices denoting the individual customer i in (8.17), using Bayes' theorem and Equations (8.5), (8.6) and (8.8), yields the sample likelihood function[24]:

$$
\begin{aligned}
&L(\lambda,\beta,r,\alpha,\hat{s},\tau_\beta,\tau_s,a,b|X,t,T)=\\
&\quad\left[\prod_{i=1}^N L_i(\lambda_i,\beta_i,a,b|X_{ij}=x_{ij},t_i,T,s_k=\hat{s_k}-\frac{1}{K}\sum\hat{s_k})\right]\left[\prod_{k=1}^K Normal(\hat{s}_k|0,\tau_s)\right]\\
&\quad\left[\prod_{i=1}^N Normal(\beta_i|1,\tau_\beta)Gamma(\lambda_i|r,\alpha)\right]
\end{aligned}
\tag{8.18}
$$

The joint posterior distribution results by substituting a and b, and multiplying the sample likelihood (8.18) with the priors (8.10)-(8.13):

$$
\begin{aligned}
&p(\lambda,\beta,r,\alpha,\hat{s},\tau_\beta,\tau_s,\phi_\mu,\phi_c|X,t,T)\propto\\
&\qquad L(\lambda,\beta,r,\alpha,\hat{s},\tau_\beta,\tau_s,a=\phi_\mu\phi_c,b=\phi_c(1-\phi_\mu)|X,t,T)\\
&\qquad Gamma(\tau_\beta|\varepsilon,\varepsilon)Gamma(\tau_s|\varepsilon,\varepsilon)Gamma(r|\varepsilon,\varepsilon)\\
&\qquad Gamma(\alpha|\varepsilon,\varepsilon)Pareto(\phi_c|\varepsilon_{min},\varepsilon_\alpha)Uniform(\phi_\mu|0,1)
\end{aligned}
\tag{8.19}
$$

The specification of the HMC simulation, using the no-U-turn sampler is based on Equation (8.19). My program code to implement the procedure is described in

[24] Here λ, β, t and s denote vector quantities, with individual elements λ_i, β_i and t_i for $i\in\{1..N\}$ and s_k for $k\in\{1..K\}$. X is the observed $N\times T$ matrix of purchase frequencies.

Appendix C.1, including hierarchical centering, variable definitions, transformations and generating forecasts.

8.2.2 Derivation of Expected Purchase Levels and $P(alive)$

In order to use the model for the prediction of individual customers' future transactions in time period $T+t_f$, I first derive the probability of $X_{T+t_F} = x_{T+t_f}$ conditional on the customer being alive in time period $T+1$. Here x_{T+t_f} is number of transactions in time period $T+t_f$. Then for $x > 0$[25]:

$$P(X_{T+tf} = x|\tau > T, \lambda, \beta, s, p, T, t_f, x > 0) = (1-p)^{t_f - 1} Poisson(x|\lambda e^{\beta s_{(t_f+T)}}) \tag{8.20}$$

The equation can be interpreted as the probability of x purchases in time-period $T+t_f$, conditional on the customer being alive in time period $T+1$ and surviving another $t_f - 1$ periods. Taking the expectation in x and marginalizing p yields the expected number of purchases in time period $T+t_f$ conditional on the customer being alive in time period $T+1$ in the desired form:

$$\begin{aligned} &E[X_{T+t_f}|\tau > T, \lambda, \beta, s, a, b, T, t_f] \\ &\quad = \sum_{x=1}^{\infty} \Big[\int_0^1 (1-p)^{t_f-1} Poisson(x|\lambda e^{\beta s_{(t_f+T)}}) Beta(p|a,b)\,\mathrm{d}p \Big]\, x \\ &\quad = e^{\beta s_{(t_f+T)}} \lambda B(a, b+t_f-1)/B(a,b) \end{aligned} \tag{8.21}$$

In the tradition of Schmittlein, Morrison, and Colombo (1987) I derive the individual probability $P(alive) = P(\tau > T)$ conditional on the observed data using the

25 For $x = 0$ the equation only extends to active customers making zero purchases. This implies that potential future zero purchase time periods, due to earlier drop-out, are not counted. As I work towards the expectation this is not relevant here, but I will revisit that issue when deriving $P(Zero_F)$.

law of total probability and Equation (8.15), setting:

$$A = \sum_{\tau=T+1}^{\infty} \Big[SGeometric(\tau|p) \prod_{j=1}^{T} Poisson(x_j|\lambda e^{\beta s_j}) \Big] \quad (8.22)$$

$$B = \sum_{\tau=t}^{T} \Big[SGeometric(\tau|p) \prod_{j=1}^{\tau} Poisson(x_j|\lambda e^{\beta s_j}) \Big] \quad (8.23)$$

Where B is the sum of probabilities that the customer drops out at $t \leq \tau \leq T$ and A is the sum of probabilities that the customer drops-out at later time period starting with $T+1$, both conditional on $(\lambda, \beta, s, p, X, t, T)$[26] . Then:

$$\begin{aligned} P(\tau > T|\lambda, \beta, s, p, t, T) \\ &= \frac{A}{A+B} = \frac{1}{1+B/A} \\ &= \frac{1}{1 + \sum_{\tau=t}^{T} e^{\lambda(\sum_{j=1}^{T} e^{\beta s_j} - \sum_{j=1}^{\tau} e^{\beta s_j})} p(1-p)^{-T+\tau-1}} \end{aligned} \quad (8.24)$$

Equation (8.24) results by applying the same transformations as in going from (8.15) to (8.16). Now integrating out p yields:

$$\begin{aligned} P(\tau > T|\lambda, \beta, s, a, b, t, T) \\ &= \int_0^1 \frac{Beta(p|a,b)}{1 + \sum_{\tau=t}^{T} e^{\lambda(\sum_{j=1}^{T} e^{\beta s_j} - \sum_{j=1}^{\tau} e^{\beta s_j})} p(1-p)^{-T+\tau-1}} \mathrm{d}p \\ &= \frac{1}{1 + \sum_{\tau=t}^{T} e^{\lambda(\sum_{j=1}^{T} e^{\beta s_j} - \sum_{j=1}^{\tau} e^{\beta s_j})} B(a+1, b+\tau-1)/B(a, b+T)} \end{aligned} \quad (8.25)$$

26 The conditioning on X can be dropped, because of the memoryless property of the Poisson and shifted geometric distributions.

With (8.25) it is now possible to remove the conditioning on the customer being alive in (8.21) and get:

$$
\begin{aligned}
&E[X_{T+t_f}|\lambda,\beta,s,a,b,t,T,t_f] \\
&\quad = E[X_{T+t_f}|\tau > T,\lambda,\beta,s,a,b,T,t_f]\,P(\tau > T|\lambda,\beta,s,a,b,t,T) \\
&\quad = \frac{e^{\beta s_{(t_f+T)}}\lambda B(a,b+t_f-1)/B(a,b)}{1+\sum_{\tau=t}^{T} e^{\lambda(\sum_{j=1}^{T} e^{\beta s_j}-\sum_{j=1}^{\tau} e^{\beta s_j})} B(a+1,b+\tau-1)/B(a,b+T)}
\end{aligned} \tag{8.26}
$$

Expectation (8.26) will be used during the HMC simulation to generate the predictions for individual future purchase levels in time periods $T+t_f$ with $t_f \in \{1..F\}$. Here F denotes the forecast horizon, the number of time periods starting with the first unobserved time period $T+1$.

8.3 A New Measure for Customer Inactivity: $P(Zero_F)$ as an Alternative to $P(alive)$

8.3.1 Conceptual Background of $P(Zero_F)$

Ever since Schmittlein, Morrison, and Colombo (1987) introduced the $P(alive)$ metric, it has been common practice for researchers to derive this metric for models that contain a form of drop-out process (Fader and Hardie 2009; Fader, Hardie, and Shang 2010). $P(alive)$ has been used for, e.g., optimizing customer reactivation campaigns (Ma, Tan, and Shu 2015), validating and benchmarking models (Batislam, Denizel, and Filiztekin 2007; Wübben and von Wangenheim 2008), customer portfolio management (Sackmann, Kundisch, and Ruch 2010), customer value analysis (Ho, Park, and Zhou 2006), and examining the effect of modes of acquisition and retention on customer lifetime (Steffes, Murthi, and Rao 2008).

At first glance $P(alive)$'s use seems attractive, because typically a customer relationship manager wants to decide at a point in time T whether a customer i is still active or not. This information is helpful to allocate marketing resources, to target customers for reactivation, and to calculate customer lifetime value and customer equity. A metric that predicts individual customer inactivity is especially appealing in noncontractual settings, where one cannot observe customer defection directly through expiration or cancellation of an ongoing contract.

But does $P(alive)$ really provide meaningful information? All it yields is information about the latent state of the customer. Would a CRM analyst not rather prefer a measure for a tangible outcome that he can observe? For example, if an analyst asks "Will this customer make any purchases in the next quarter?", $P(alive)$ is not the correct measure. Or if a marketing manager wants to avoid offering incentives to customers that are likely to make no purchases in the next year — again, $1 - P(alive)$ is not the correct measure. Remarkably, these two cases are exemplary for situations $P(alive)$ is currently used for.

The reason for this discrepancy is that $P(alive)$ pertains to an infinite time horizon. For example, $1 - P(alive)$ only yields a sound theoretical probability to the question "How likely is it that this customer never purchases again?" Conversely, $P(alive)$ would provide the theoretical correct measure to the question "Will this customer make another purchase, ever?" It seems odd that managers would use a such a measure when they are interested in the foreseeable future or a certain planning horizon. In most managerial decision situations the forecast horizon ranges from the next quarter, the next year to maybe the next 5 or 10 years at most. This very issue is raised by Wübben and von Wangenheim (2008, p. 91), who, while discussing $P(alive)$ state: "... it is of hardly any interest to managers whether a customer purchases after the planning horizon."

Even more troubling for researchers is that predictions based on $P(alive)$ cannot be empirically validated. True customer inactivity cannot be observed with finite length hold-out periods. Currently, researchers use long periods of inactivity as a proxy for true customer inactivity in out-of-sample validation. However, this practice might introduce systematic bias and lead to inaccurate predictions. In that respect $P(alive)$ is of limited diagnostic value when viewed by itself (Fader, Hardie, and Shang 2010).

For the HSMDO, I provide a solution to the aforementioned challenges and propose an alternative measure $P(Zero_F)$ that is a theoretically sound probabilistic measure for customer inactivity in finite time horizons, provides managerially relevant information on observable outcomes, and is flexible because the desired time horizon F can be chosen freely.

The main idea behind deriving $P(Zero_F)$ is that the observation of zero purchases in the forecast horizon has one of three reasons: (1) true customer inactivity due to drop-out before the hold-out period, the case covered by $1-P(alive)$; (2) the chance event that the individual customer, independent of his purchase rate, makes no purchases in the hold-out period without dropping out; and (3) the customer is alive at the beginning of the hold-out period, but drops out during he hold-out period and makes zero purchases until drop-out. Cases (2) and (3) are not covered by $1-P(alive)$. This implies that $1-P(alive)$ systematically underestimates the probability of seeing zero purchases in the forecast period.

$P(Zero_F)$ takes into account all three cases stated above. $P(Zero_F)$ is the probability that a customer will make no purchases in the forecast period (from $T+1$ to $T+F$) or alternatively $1-P(Zero_F)$ is the probability that the customer will make at least one purchase in the next F time periods. $P(Zero_F)$ provides the theoretically "correct"[27] answers to the managerial questions I raised at the beginning of this section. For example, $1-P(Zero_3)$ would be the probabilistic answer[28] to the question: "Will the customer make any purchases in the next quarter?" $P(Zero_{12})$ would be the probabilistic answer to the question: "How likely is it that this customer makes no purchases in the next year?"

In the next section I will derive $P(Zero_F)$ for the HSMDO mathematically and in Section 8.4.2.2 I will test if $P(Zero_F)$'s theoretical properties translate into improved forecast accuracy of customer inactivity.

27 Within the confines of the probabilistic model.

28 Assuming monthly time units.

8.3.2 Derivation of $P(Zero_F)$ for the HSMDO

In order to derive $P(Zero_F)$ I use the mathematical framework and variable definitions as in Section 8.2. Specifically, x_j denotes the number of purchases the customer made in time period j, T is the length of the observation period, t denotes the time period of the last observed purchase, s is the vector of the seasonal components, p is the individual drop-out rate, λ is the individual purchase rate and β is the individual seasonality. The customer specific index i is dropped for brevity and $s_j = s_{k(j)}$. Using the individual level likelihood from (8.16) conditional on $(X_j = x_j, t, T, s)$ and setting:

$$A = (1-p)^T \prod_{j=1}^{T} Poisson(x_j|\lambda e^{\beta s_j}) \tag{8.27}$$

$$B = \sum_{\tau=t}^{T} (1-p)^{\tau-1} p \prod_{j=1}^{\tau} Poisson(x_j|\lambda e^{\beta s_j}) \tag{8.28}$$

$$C = (1-p)^F \prod_{j=T+1}^{T+F} Poisson(0|\lambda e^{\beta s_j}) \tag{8.29}$$

$$D = \sum_{\tau=T+1}^{T+F} (1-p)^{\tau-1} p \prod_{j=1}^{T} Poisson(x_j|\lambda e^{\beta s_j}) \prod_{j=T+1}^{\tau} Poisson(0|\lambda e^{\beta s_j}) \tag{8.30}$$

The term $A+B$ is the individual level likelihood from (8.16), while $A \cdot C$ constrains potential purchases in the forecast horizon F to zero, for customers, who remain active throughout the forecast horizon $\tau > (T+F)$. $B+D$ constrains purchases in the forecast horizon to zero for customers, who drop-out during the forecast horizon or earlier. Then the probability of observing zero purchases in forecast horizon F, due to drop-out and/or no purchases while active, is:

$$P(Zero_F|\lambda, \beta, p, X_j = x_j, t, T, s) = \frac{A \cdot C + B + D}{A+B} \tag{8.31}$$

$$= \frac{C + B/A + D/A}{1 + B/A} \tag{8.32}$$

Canceling common factors, substituting and simplifying products to exponentiated sums yields:

$$
\begin{aligned}
&P(Zero_F|\lambda,\beta,p,X_j=x_j,t,T,s)\\
&\quad=\frac{1}{1+B/A}\Big[B/A+(1-p)^F\prod_{j=T+1}^{T+F}Poisson(0|\lambda e^{\beta s_j})\\
&\qquad\qquad\qquad+\sum_{\tau=T+1}^{T+F}(1-p)^{\tau-1-T}p\prod_{j=T+1}^{\tau}Poisson(0|\lambda e^{\beta s_j})\Big]\\
&\quad=\frac{1}{1+B/A}\Big[B/A+(1-p)^F e^{-\lambda\sum_{j=1}^{F}e^{\beta s_{j+T}}}+\sum_{\tau=1}^{F}(1-p)^{\tau-1}pe^{-\lambda\sum_{j=1}^{\tau}e^{\beta s_{j+T}}}\Big]\\
&\text{with}\quad B/A=\sum_{\tau=t}^{T}p(1-p)^{-T+\tau-1}e^{\lambda(\sum_{j=1}^{T}e^{\beta s_j}-\sum_{j=1}^{\tau}e^{\beta s_j})}
\end{aligned}
\tag{8.33}
$$

It is noteworthy that the term $1/(1+B/A)$ is equivalent to $P(\tau>T)=P(alive)$. This allows me to rearrange terms and restate the equation in an alternative form[29]:

$$
\begin{aligned}
&P(Zero_F)\\
&\quad=\frac{B/A}{1+B/A}+\frac{1}{1+B/A}\Big[(1-p)^F e^{-\lambda\sum_{j=1}^{F}e^{\beta s_{j+T}}}+\sum_{\tau=1}^{F}(1-p)^{\tau-1}pe^{-\lambda\sum_{j=1}^{\tau}e^{\beta s_{j+T}}}\Big]\\
&\quad=P(\tau\le T)+P(\tau>T)\Big[(1-p)^F e^{-\lambda\sum_{j=1}^{F}e^{\beta s_{j+T}}}+\sum_{\tau=1}^{F}(1-p)^{\tau-1}pe^{-\lambda\sum_{j=1}^{\tau}e^{\beta s_{j+T}}}\Big]
\end{aligned}
\tag{8.34}
$$

Equation (8.34) can intuitively be interpreted. The probability for observing no purchases in $T+1$ to $T+F$ is (1) the probability that the customer dropped out before $T+1$. If (2) he did not drop-out then, he might stay active another F periods $(1-p)^F$ making zero purchases in the meantime, or (3) he might drop-out in any of the time periods $T+1, T+2, ..., T+F$, making zero purchases before drop-out.

Finally, in order to integrate out p I will use Bayes' Theorem to reduce the task of integrating (8.33) to integrating over a sum. I use the individual level likelihoods

[29] Dropping the conditioning for brevity.

(8.16) and (8.17)[30] to yield the following simplification:

$$
\begin{aligned}
&P(Zero_F|\lambda,\beta,a,b,X_j,t,T,s)\\
&=\int_0^1 P(Zero_F|\lambda,\beta,p,X_j,t,T,s)Beta(p|a.b)\,\mathrm{d}p\\
&=\int_0^1 \frac{A\cdot C+B+D}{L(\lambda,\beta,p|X_j,t,T,s)}\frac{L(\lambda,\beta,p|X_j,t,T,s)Beta(p|a,b)}{L(\lambda,\beta,a,b|X_j,t,T,s)}\,\mathrm{d}p\\
&=\int_0^1 \frac{(A\cdot C+B+D)Beta(p|a,b)}{L(\lambda,\beta,a,b|X_j,t,T,s)}\,\mathrm{d}p
\end{aligned}
\tag{8.35}
$$

Setting:

$$
\begin{aligned}
U&=\int_0^1 (A\cdot C)\,Beta(p|a.b)\,\mathrm{d}p\\
&=\frac{B(a,b+T+F)}{B(a,b)}\left(\prod_{j=T+1}^{T+F} e^{-\lambda e^{\beta s_j}}\right)\prod_{j=1}^{T}\frac{\lambda^{x_j}e^{\beta s_j x_j}e^{-\lambda e^{\beta s_j}}}{\Gamma\left(x_j+1\right)}\\
&=\frac{B(a,b+T+F)}{B(a,b)\prod_{j=1}^{T}\Gamma\left(x_j+1\right)}\frac{\lambda^{\sum_{j=1}^{T}x_j}e^{\beta\sum_{j=1}^{T}s_jx_j}}{e^{\lambda\sum_{j=T+1}^{T+F}e^{\beta s_j}}e^{\lambda\sum_{j=1}^{T}e^{\beta s_j}}}
\end{aligned}
\tag{8.36}
$$

$$
\begin{aligned}
V&=\int_0^1 B\cdot Beta(p|a.b)\,\mathrm{d}p\\
&=\sum_{\tau=t}^{T}\frac{B(a+1,b+\tau-1)}{B(a,b)}\prod_{j=1}^{\tau}\frac{\lambda^{x_j}e^{\beta s_j x_j}e^{-\lambda e^{\beta s_j}}}{\Gamma\left(x_j+1\right)}\\
&=\sum_{\tau=t}^{T}\frac{B(a+1,b+\tau-1)}{B(a,b)\prod_{j=1}^{\tau}\Gamma\left(x_j+1\right)}\lambda^{\sum_{j=1}^{\tau}x_j}e^{\beta\sum_{j=1}^{\tau}s_jx_j}e^{-\lambda\sum_{j=1}^{\tau}e^{\beta s_j}}
\end{aligned}
\tag{8.37}
$$

$$
\begin{aligned}
W&=\int_0^1 D\cdot Beta(p|a.b)\,\mathrm{d}p\\
&=\sum_{\tau=T+1}^{T+F}\frac{B(a+1,b+\tau-1)}{B(a,b)}\left(\prod_{j=T+1}^{\tau} e^{-\lambda e^{\beta s_j}}\right)\prod_{j=1}^{T}\frac{\lambda^{x_j}e^{\beta s_j x_j}e^{-\lambda e^{\beta s_j}}}{\Gamma\left(x_j+1\right)}\\
&=\sum_{\tau=T+1}^{T+F}\frac{B(a+1,b+\tau-1)}{B(a,b)\prod_{j=1}^{T}\Gamma\left(x_j+1\right)}\frac{\lambda^{\sum_{j=1}^{T}x_j}e^{\beta\sum_{j=1}^{T}s_jx_j}}{e^{\lambda\sum_{j=T+1}^{\tau}e^{\beta s_j}}e^{\lambda\sum_{j=1}^{T}e^{\beta s_j}}}
\end{aligned}
\tag{8.38}
$$

30 Here I restore the factor $\prod_{j=1}^{t}(1/\Gamma(x_j+1))$.

Substituting and canceling terms results in:

$$P(Zero_F|\lambda,\beta,a,b,X_j,t,T,s) = \frac{U+V+W}{L(\lambda,\beta,a,b|X_j,t,T,s)}$$
$$= \Big[\frac{B(a,b+T+F)}{e^{\lambda\sum_{j=T+1}^{T+F}e^{\beta s_j}}e^{\lambda\sum_{j=1}^{T}e^{\beta s_j}}} + \sum_{\tau=t}^{T}\frac{B(a+1,b+\tau-1)}{e^{\lambda\sum_{j=1}^{\tau}e^{\beta s_j}}} + \sum_{\tau=T+1}^{T+F}\frac{B(a+1,b+\tau-1)}{e^{\lambda\sum_{j=T+1}^{\tau}e^{\beta s_j}}e^{\lambda\sum_{j=1}^{T}e^{\beta s_j}}}\Big]$$
$$\cdot\Big[B(a,b+T)e^{-\lambda\sum_{j=1}^{T}e^{\beta s_j}} + \sum_{\tau=t}^{T}B(1+a,b+\tau-1)e^{-\lambda\sum_{j=1}^{\tau}e^{\beta s_j}}\Big]^{-1} \tag{8.39}$$

Now dividing by $B(a,b+T)e^{-\lambda\sum_{j=1}^{T}e^{\beta s_j}}$ and substituting R yields the desired form:

$$R = \sum_{\tau=t}^{T}\frac{B(1+a,b+\tau-1)}{B(a,b+T)}e^{\lambda(\sum_{j=1}^{T}e^{\beta s_j}-\sum_{j=1}^{\tau}e^{\beta s_j})} \tag{8.40}$$

$$P(Zero_F|\lambda,\beta,a,b,X_j,t,T,s)$$
$$=\Big[1+R\Big]^{-1}\Big[R+\frac{B(a,b+T+F)}{e^{\lambda\sum_{j=T+1}^{T+F}e^{\beta s_j}}B(a,b+T)} + \sum_{\tau=T+1}^{T+F}\frac{B(a+1,b+\tau-1)}{e^{\lambda\sum_{j=T+1}^{\tau}e^{\beta s_j}}B(a,b+T)}\Big] \tag{8.41}$$

Equation 8.41 is used in the HMC simulation to compute $P(Zero_F)$ in every iteration.

8.4 Data Analysis: Parameter Estimation and Prediction

8.4.1 Convergence and Marginal Posterior Estimates

I obtained the posterior distributions for the model's parameters by Hybrid/Hamilton Monte Carlo Sampling (Brooks et al. 2011, pp. 113-160), specifically the no-U-turn sampler (Homan and Gelman 2014). For this model the no-U-turn sampler substantially increases the effective sample size relative to computational time compared to Gibbs sampling that I used for the DMPT and HSM. I present my source code and additional information about the procedure in Appendix C.1.

The parameter estimation results are based on HMC runs with four chains starting from dispersed initial values, 20,000 iterations and a burn-in phase of 10,000 samples. I performed thinning by taking every fifth sample after the burn-in period to save memory. Therefore, the marginal posterior distributions are based on $(10,000/5) \times 4 = 8,000$ samples.

The HMC procedure iteratively converges towards the true posterior distribution of the parameter space and after convergence simulates the true posterior distribution. I assessed approximate convergence of the algorithm by examining auto-correlation and effective sample size; visually inspecting the trace plots of the posterior samples; evaluating the Gelman, Brooks and Rubin statistic; and by running multiple simulations from dispersed initial values.

Auto-Correlation and Effective Sample Size

The samples obtained via HMC for all three datasets show virtually no auto-correlation after the burn-in period (10,000 iterations). The auto-correlation plots for the main parameters of the model and seasonal components are shown in Figures C.1 and C.3 (Apparel), Figures C.5 and C.7 (DIY), and Figures C.9 and C.11 (CDNOW) in the Appendix. The lowest two effective sample sizes for the DIY HMC simulation are 6,876 for τ_s and 7,089 for τ_β, while the two lowest effective sample sizes for the apparel samples are 6,877 for τ_β and 7,318 for s_{12}. For the CDNOW dataset, the procedure yielded the fewest effective samples for r (4,036) and α (4,221). Typically effective sample sizes were near their theoretical maximum of 8,000 samples. Thus, auto-correlation is not an issue with the obtained effective sample sizes (Gelman et al. 2003, p. 298).

Inspection of Trace Plots

The trace plots for all three datasets exhibit no visible "trends" and sufficient mixing of all four Markov chains both individually and overall. The trace plots for the main parameters of the model and seasonal components are shown in Figures C.2 and C.4 (Apparel), Figures C.6 and C.8 (DIY), and Figures C.10 and C.12 (CDNOW) in the Appendix. The visual inspection of the trace plots indicates that convergence was reached and that the posterior distribution is adequately represented (see Section 4.3).

BGR Statistic

I validated Markov chain convergence by computing the Brooks-Gelman-Rubin statistic for all parameters over all four chains. In all cases, the BGR ratio $\hat{R}_c$ was approximately 1.0 and stayed well below 1.1 close to the ideal 1.0, which indicates that the chains have converged to the true posterior distribution (Congdon 2003; Cowles and Carlin 1996; Gelman and Rubin 1992). In addition, I repeatedly simulated the posterior distribution using multi-chain runs with dispersed initial values that in all cases converged to virtually identical parameter values.

Parameter Estimates and Marginal Posterior Densities

Summary statistics for the parameter's marginal posterior distributions including means, standard deviations, medians, and credibility intervals are compiled in Tables 8.1 (DIY), 8.2 (Apparel), and 8.3 (CDNOW). The tables include all main model parameters $r, \alpha, \tau_s, \tau_\beta, a$, and b, as well as all seasonal components $s_1 - s_{12}$. Additionally, the first and last two of the individual level parameters β_i, λ_i, $PA_i = P_i(alive)$, and $PZF_i = P_i(Zero_F)$ are shown. The estimates for $P_i(alive)$ and $P_i(Zero_F)$ are computed at each iteration of the HMC simulation via Equations (8.41) and (8.25). The full marginal posterior densities of the main model parameters and seasonal components are depicted in Figures 8.7-8.12.

Table 8.1: HSMDO/DIY Marginal Posterior Distribution Summaries for HMC Simulation after 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

Summary of Marginal Posterior Distributions						
node	*mean*	*sd*	*MC error*	*2.5%*	*median*	*97.5%*
r	2.1803	0.0823	0.0010	2.0245	2.1778	2.3486
alpha	0.7796	0.0330	0.0004	0.7169	0.7789	0.8464
s[1]	0.0681	0.0072	0.0001	0.0542	0.0680	0.0821
s[2]	0.0394	0.0082	0.0001	0.0226	0.0396	0.0549
s[3]	0.0209	0.0076	0.0001	0.0056	0.0210	0.0355
s[4]	−0.0045	0.0095	0.0001	−0.0231	−0.0044	0.0137
s[5]	−0.0156	0.0104	0.0001	−0.0361	−0.0156	0.0046
s[6]	−0.0953	0.0106	0.0001	−0.1161	−0.0952	−0.0750
s[7]	−0.2358	0.0140	0.0002	−0.2633	−0.2356	−0.2085
s[8]	−0.1758	0.0127	0.0001	−0.2013	−0.1755	−0.1515
s[9]	0.0356	0.0092	0.0001	0.0177	0.0356	0.0538
s[10]	0.1464	0.0086	0.0001	0.1294	0.1465	0.1630
s[11]	0.1286	0.0078	0.0001	0.1135	0.1286	0.1438
s[12]	0.0880	0.0077	0.0001	0.0731	0.0880	0.1032
tau_s	72.2461	31.2178	0.3765	24.3840	67.7148	144.7251
tau_beta	0.4584	0.0456	0.0005	0.3748	0.4561	0.5523
a	1.0085	3.3563	0.0379	0.0980	0.4312	5.0562
b	672.1070	2232.5177	25.2373	63.0831	286.2169	3423.5955
lambda[1]	7.5578	0.5321	0.0060	6.5502	7.5449	8.6265
lambda[2]	2.6883	0.3127	0.0035	2.1112	2.6804	3.3320
⋮	⋮	⋮	⋮	⋮	⋮	⋮
lambda[1328]	1.3751	0.2203	0.0025	0.9713	1.3619	1.8395
lambda[1329]	4.4228	0.3981	0.0045	3.7004	4.4056	5.2614
beta[1]	2.5930	0.6660	0.0075	1.3029	2.5799	3.9609
beta[2]	1.9810	0.9348	0.0105	0.1586	1.9760	3.8561
⋮	⋮	⋮	⋮	⋮	⋮	⋮
beta[1328]	1.5524	1.1046	0.0124	−0.6111	1.5505	3.7900
beta[1329]	0.8533	0.7806	0.0087	−0.6425	0.8547	2.3994
PA[1]	0.9986	0.0003	0.0000	0.9981	0.9986	0.9991
PA[2]	0.9986	0.0003	0.0000	0.9981	0.9986	0.9991
⋮	⋮	⋮	⋮	⋮	⋮	⋮
PA[1328]	0.9986	0.0003	0.0000	0.9981	0.9986	0.9991
PA[1329]	0.9986	0.0003	0.0000	0.9981	0.9986	0.9991
PZF[1]	0.0014	0.0003	0.0000	0.0009	0.0014	0.0019
PZF[2]	0.0017	0.0004	0.0000	0.0011	0.0016	0.0028
⋮	⋮	⋮	⋮	⋮	⋮	⋮
PZF[1328]	0.0128	0.0101	0.0001	0.0026	0.0098	0.0405
PZF[1329]	0.0014	0.0003	0.0000	0.0009	0.0014	0.0019

Table 8.2: HSMDO/Apparel Marginal Posterior Distribution Summaries for HMC Simulation after 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

Summary of Marginal Posterior Distributions						
node	*mean*	*sd*	*MC error*	*2.5%*	*median*	*97.5%*
r	1.7851	0.0842	0.0010	1.6271	1.7846	1.9516
alpha	2.3915	0.1237	0.0014	2.1590	2.3880	2.6372
s[1]	0.6991	0.0207	0.0002	0.6589	0.6992	0.7397
s[2]	−0.5257	0.0548	0.0006	−0.6360	−0.5244	−0.4204
s[3]	−0.2291	0.0400	0.0005	−0.3103	−0.2277	−0.1550
s[4]	−0.1020	0.0392	0.0004	−0.1802	−0.1017	−0.0281
s[5]	−0.0880	0.0378	0.0004	−0.1638	−0.0878	−0.0163
s[6]	0.0885	0.0299	0.0003	0.0288	0.0887	0.1454
s[7]	0.2970	0.0263	0.0003	0.2452	0.2971	0.3473
s[8]	0.1309	0.0337	0.0004	0.0633	0.1308	0.1957
s[9]	−0.7375	0.0838	0.0010	−0.9060	−0.7350	−0.5794
s[10]	0.1210	0.0391	0.0004	0.0435	0.1211	0.1977
s[11]	−0.1632	0.0493	0.0006	−0.2643	−0.1623	−0.0694
s[12]	0.5089	0.0273	0.0003	0.4560	0.5089	0.5627
tau_s	6.0677	2.6474	0.0296	2.0844	5.7092	12.3811
tau_beta	1.5511	0.1339	0.0016	1.3089	1.5445	1.8331
a	0.3886	0.7971	0.0092	0.1144	0.2388	1.5670
b	12.1896	33.3422	0.3865	1.6886	5.7570	61.2676
ambda[1]	1.5805	0.2680	0.0030	1.1021	1.5643	2.1694
lambda[2]	2.3098	0.3326	0.0037	1.6995	2.2951	3.0155
⋮	⋮	⋮	⋮	⋮	⋮	⋮
lambda[1229]	0.5379	0.1560	0.0017	0.2692	0.5223	0.8790
lambda[1230]	1.2832	0.2453	0.0027	0.8461	1.2709	1.8007
beta[1]	0.6269	0.4040	0.0045	−0.1650	0.6259	1.4206
beta[2]	0.7230	0.3562	0.0042	0.0259	0.7201	1.4170
⋮	⋮	⋮	⋮	⋮	⋮	⋮
beta[1229]	0.8029	0.5663	0.0063	−0.2932	0.8023	1.9022
beta[1230]	0.8027	0.4317	0.0050	−0.0399	0.8035	1.6619
PA[1]	0.9897	0.0035	0.0000	0.9810	0.9905	0.9945
PA[2]	0.8670	0.0569	0.0006	0.7249	0.8776	0.9448
⋮	⋮	⋮	⋮	⋮	⋮	⋮
PA[1229]	0.9309	0.0297	0.0003	0.8551	0.9371	0.9686
PA[1230]	0.7600	0.1127	0.0013	0.4868	0.7824	0.9135
PZF[1]	0.0175	0.0067	0.0001	0.0084	0.0162	0.0337
PZF[2]	0.1357	0.0569	0.0006	0.0577	0.1248	0.2776
⋮	⋮	⋮	⋮	⋮	⋮	⋮
PZF[1229]	0.1991	0.0630	0.0007	0.1100	0.1880	0.3602
PZF[1230]	0.2513	0.1083	0.0012	0.1103	0.2281	0.5169

Table 8.3: HSMDO/CDNOW Marginal Posterior Distribution Summaries for HMC Simulation after 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

Summary of Marginal Posterior Distributions						
node	*mean*	*sd*	*MC error*	*2.5%*	*median*	*97.5%*
r	6.2802	0.8147	0.0128	4.8697	6.2143	8.0867
alpha	16.4389	2.1999	0.0339	12.6043	16.2577	21.2661
s[1]	0.9469	0.0466	0.0006	0.8561	0.9466	1.0383
s[2]	−0.0187	0.0676	0.0008	−0.1513	−0.0181	0.1129
s[3]	−0.1014	0.0869	0.0010	−0.2754	−0.1001	0.0662
s[4]	−0.2244	0.0946	0.0011	−0.4156	−0.2231	−0.0456
s[5]	−0.2185	0.0967	0.0011	−0.4126	−0.2168	−0.0321
s[6]	−0.1219	0.0951	0.0011	−0.3123	−0.1204	0.0628
s[7]	−0.0442	0.0953	0.0011	−0.2351	−0.0411	0.1388
s[8]	−0.1721	0.1043	0.0012	−0.3801	−0.1693	0.0288
s[9]	−0.0726	0.1038	0.0012	−0.2806	−0.0714	0.1249
s[10]	−0.1978	0.1134	0.0013	−0.4268	−0.1952	0.0203
s[11]	0.1319	0.1016	0.0011	−0.0713	0.1318	0.3249
s[12]	0.0929	0.1069	0.0012	−0.1191	0.0943	0.2953
tau_s	9.0783	4.0115	0.0449	3.0697	8.4102	18.8614
tau_beta	739.7070	699.7777	7.8541	104.0761	524.7925	2638.4254
a	0.4979	0.0736	0.0008	0.3733	0.4934	0.6599
b	0.8208	0.2109	0.0024	0.4973	0.7905	1.3251
lambda[1]	0.3180	0.1012	0.0011	0.1545	0.3075	0.5411
lambda[2]	0.4268	0.1537	0.0017	0.1804	0.4106	0.7685
⋮	⋮	⋮	⋮	⋮	⋮	⋮
lambda[780]	0.3735	0.1424	0.0016	0.1518	0.3556	0.7007
lambda[781]	0.3707	0.1399	0.0016	0.1505	0.3524	0.6944
beta[1]	1.0017	0.0520	0.0006	0.8959	1.0011	1.1062
beta[2]	1.0027	0.0521	0.0006	0.9013	1.0016	1.1098
⋮	⋮	⋮	⋮	⋮	⋮	⋮
beta[780]	1.0000	0.0523	0.0006	0.8884	1.0002	1.1096
beta[781]	0.9996	0.0531	0.0006	0.8905	0.9990	1.1108
PA[1]	0.8078	0.0467	0.0005	0.6946	0.8164	0.8753
PA[2]	0.0043	0.0078	0.0001	0.0000	0.0013	0.0263
⋮	⋮	⋮	⋮	⋮	⋮	⋮
PA[780]	0.0071	0.0115	0.0001	0.0000	0.0028	0.0384
PA[781]	0.0070	0.0108	0.0001	0.0000	0.0029	0.0388
PZF[1]	0.5038	0.0718	0.0008	0.3936	0.4925	0.6652
PZF[2]	0.9977	0.0032	0.0000	0.9883	0.9990	1.0000
⋮	⋮	⋮	⋮	⋮	⋮	⋮
PZF[780]	0.9965	0.0041	0.0001	0.9851	0.9981	1.0000
PZF[781]	0.9965	0.0040	0.0001	0.9851	0.9981	1.0000

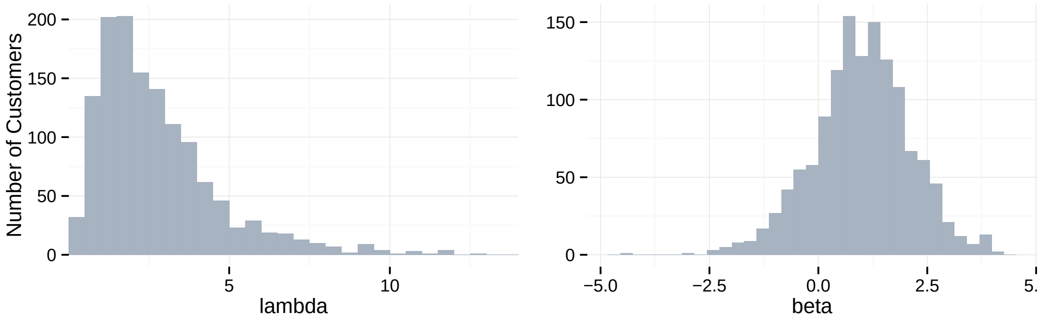

Figure 8.1: HSMDO/DIY Customer Base Histograms for Estimates of Individual Purchase Rate λ_i and Customer Seasonality β_i

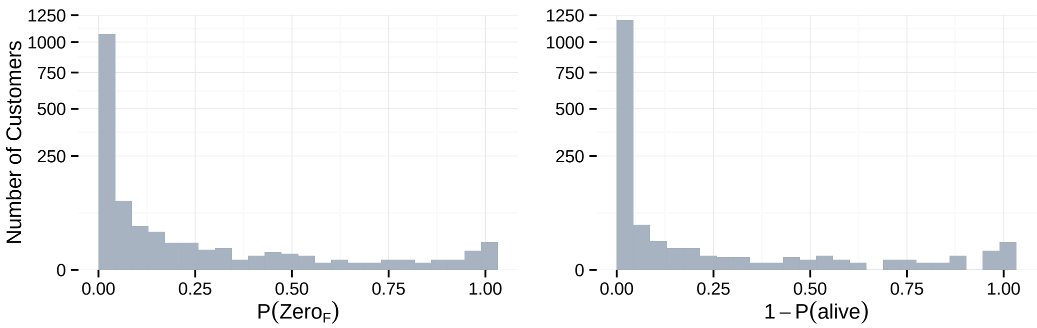

Figure 8.2: HSMDO/DIY Customer Base Histograms for Estimates of Individual Customer Inactivity $P(Zero_F)$ and $1 - P(alive)$

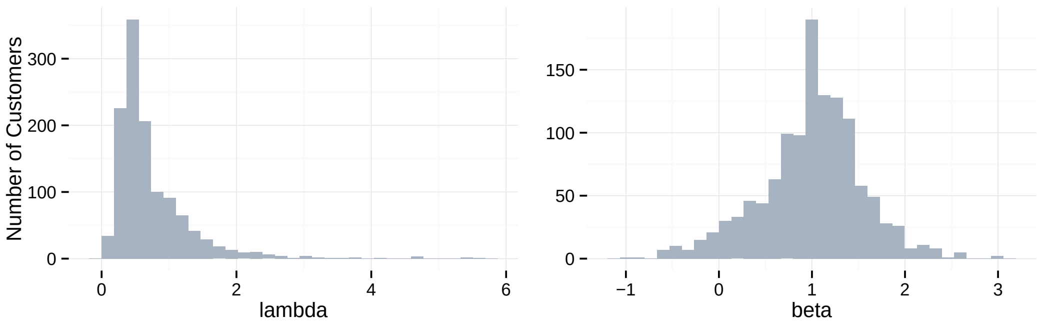

Figure 8.3: HSMDO/Apparel Customer Base Histograms for Estimates of Individual Purchase Rate λ_i and Customer Seasonality β_i

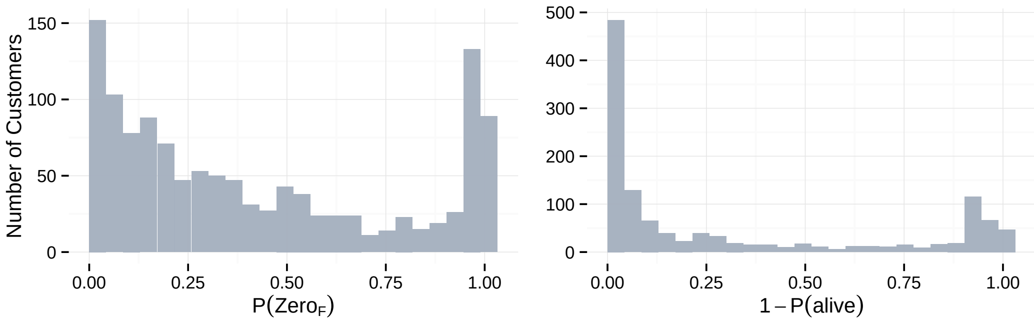

Figure 8.4: HSMDO/Apparel Customer Base Histograms for Estimates of Individual Customer Inactivity $P(Zero_F)$ and $1 - P(alive)$

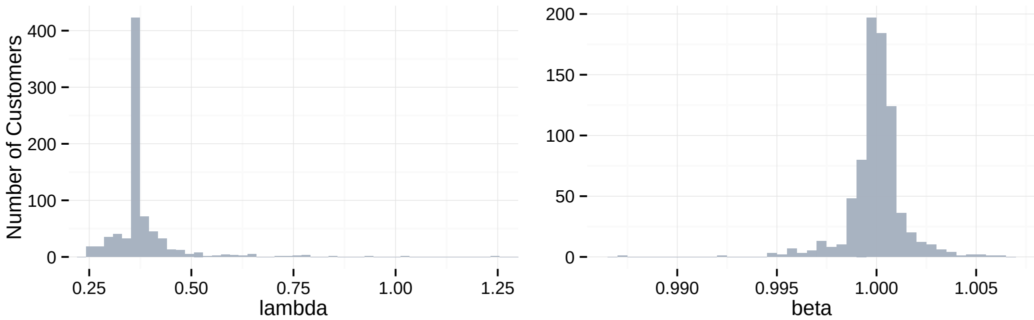

Figure 8.5: HSMDO/CDNOW Customer Base Histograms for Estimates of Individual Purchase Rate λ_i and Customer Seasonality β_i

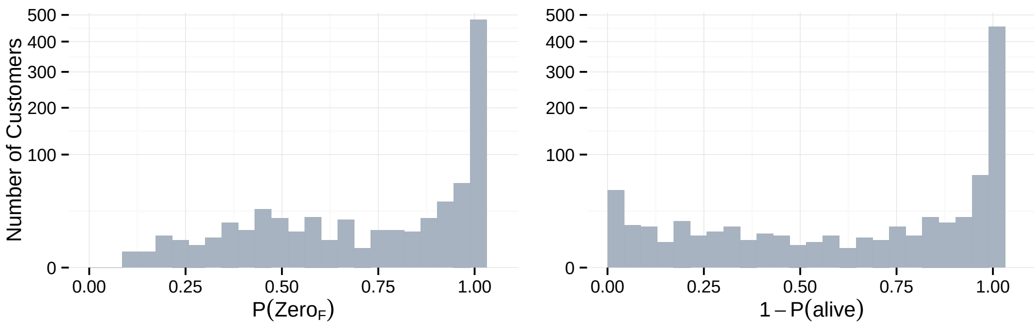

Figure 8.6: HSMDO/CDNOW Customer Base Histograms for Estimates of Individual Customer Inactivity $P(Zero_F)$ and $1 - P(alive)$

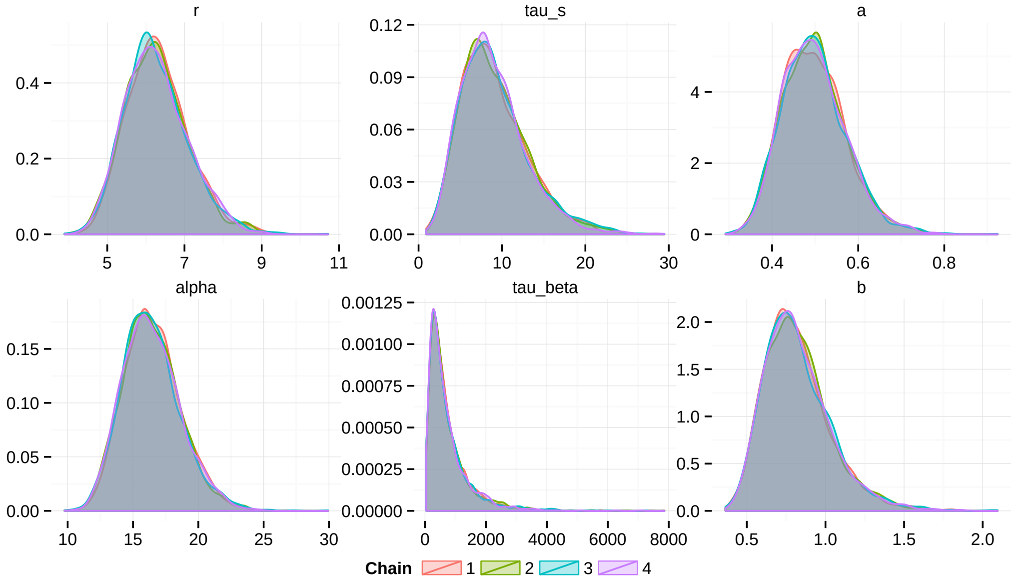

Figure 8.7: HSMDO/CDNOW Marginal Posterior Densities for r, α, τ_s, τ_β, a, and b after 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

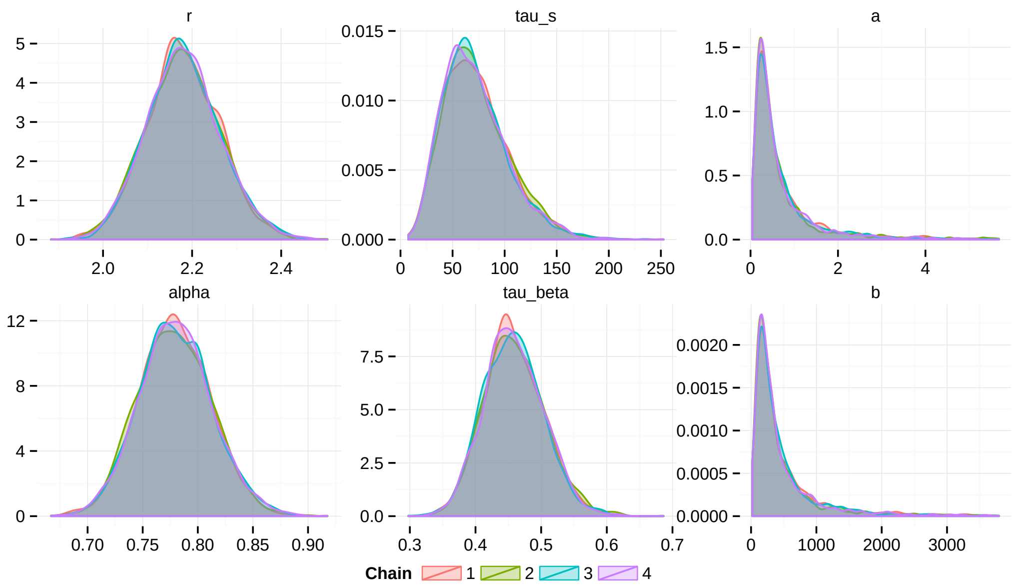

Figure 8.8: HSMDO/DIY Marginal Posterior Densities for r, α, τ_s, τ_β, a, and b after 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

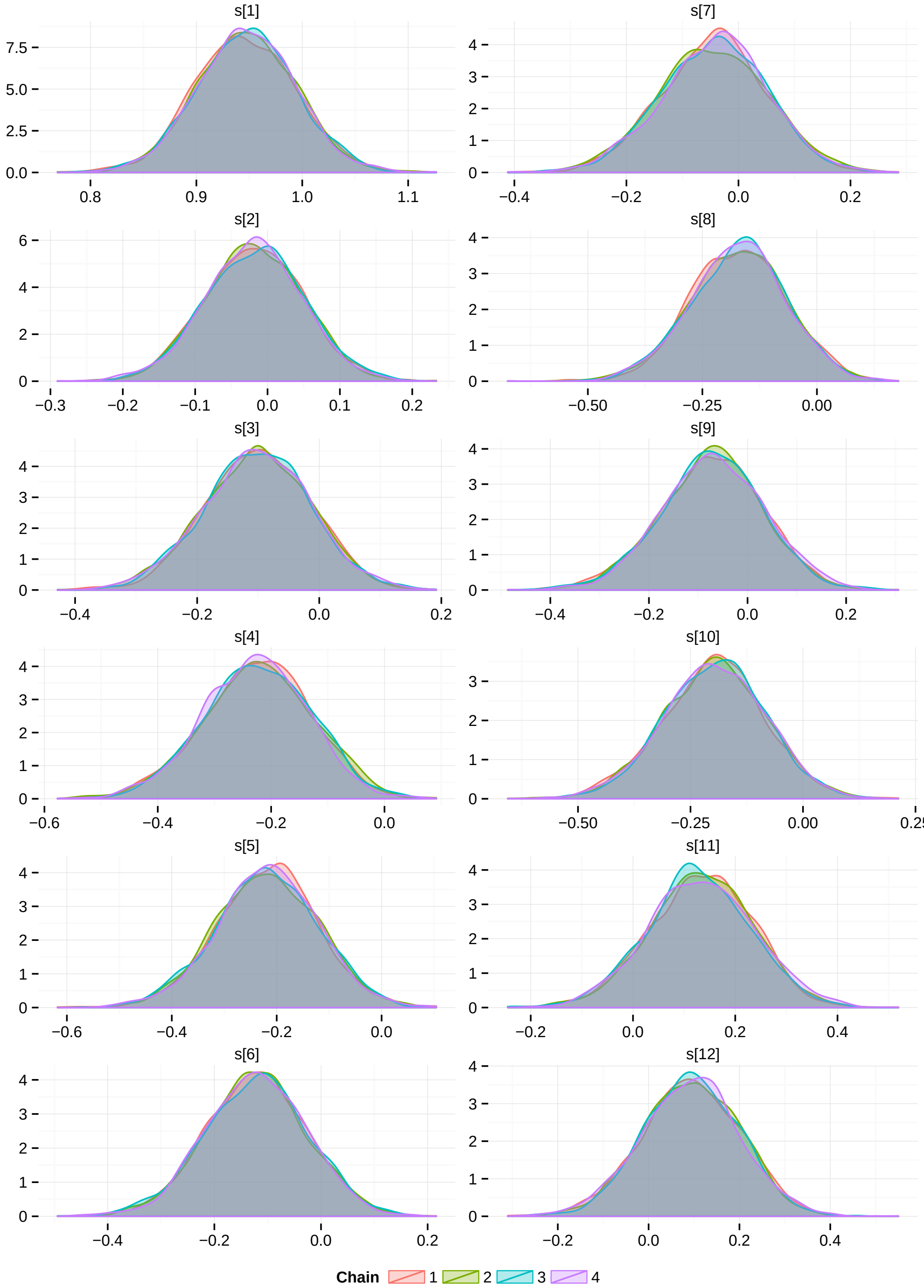

Figure 8.9: HSMDO/CDNOW Marginal Posterior Densities Seasonal Components s_k after 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

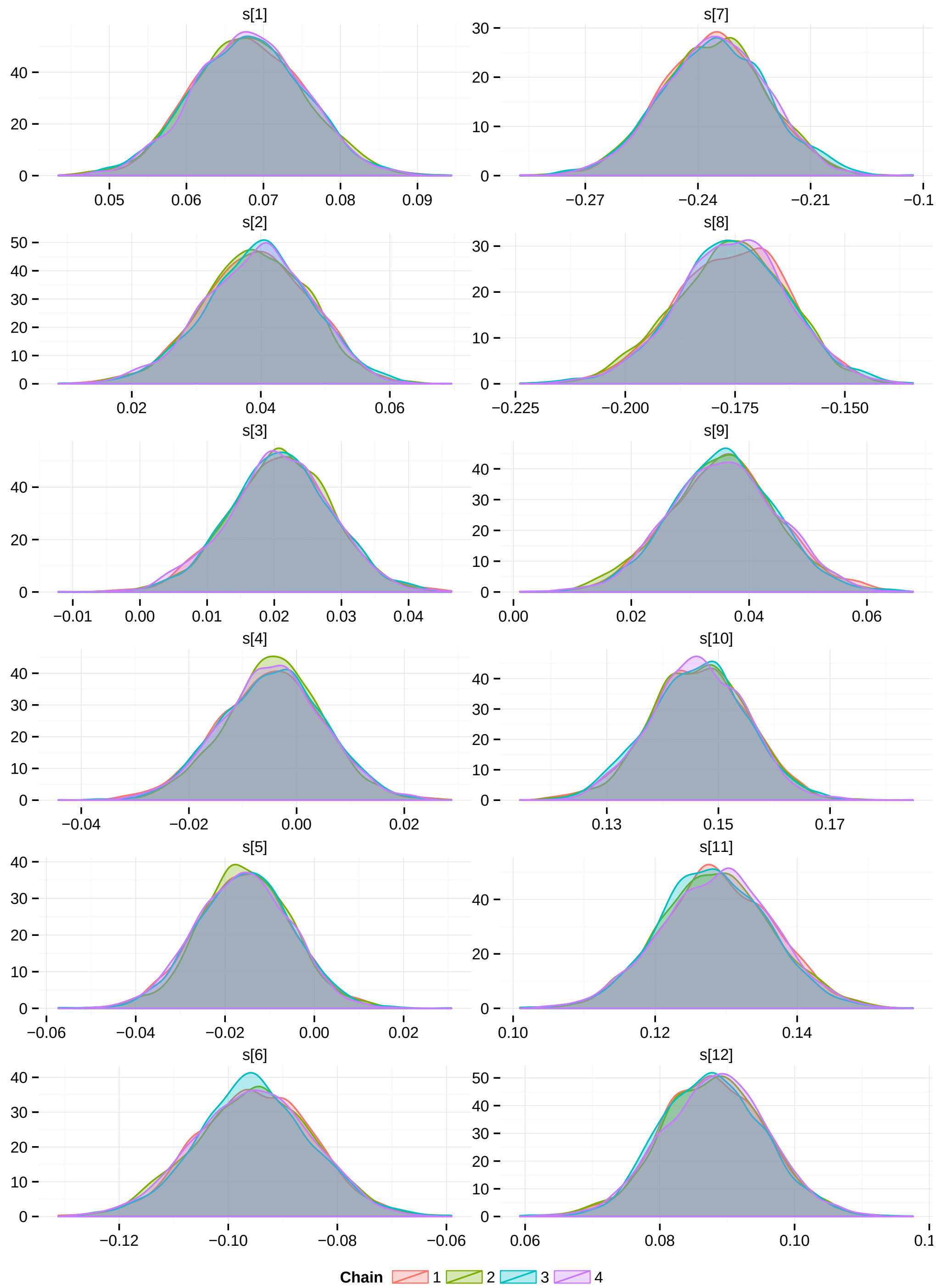

Figure 8.10: HSMDO/DIY Marginal Posterior Densities for Seasonal Components s_k after 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

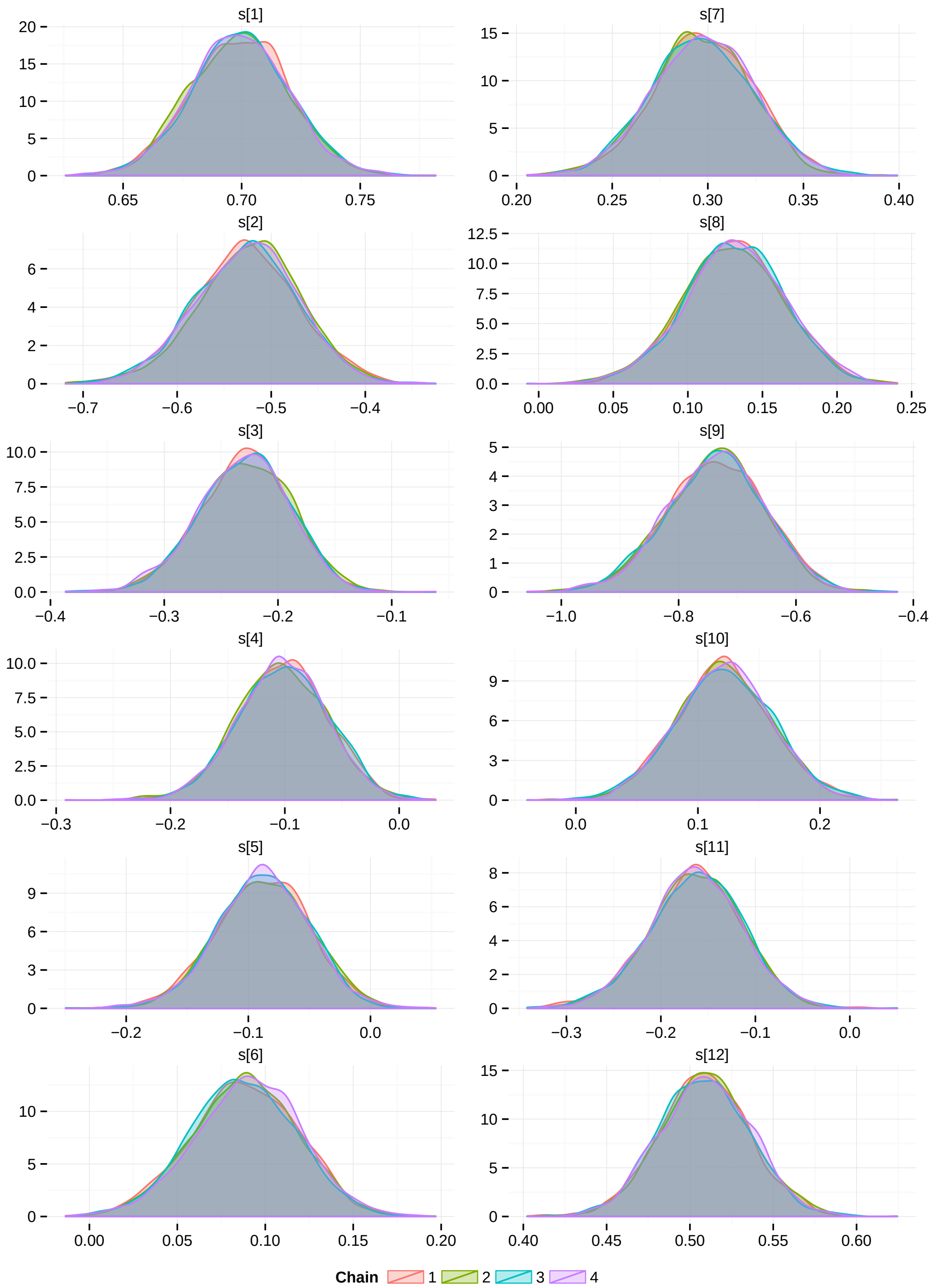

Figure 8.11: HSMDO/Apparel Marginal Posterior Densities Seasonal Components s_k after 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

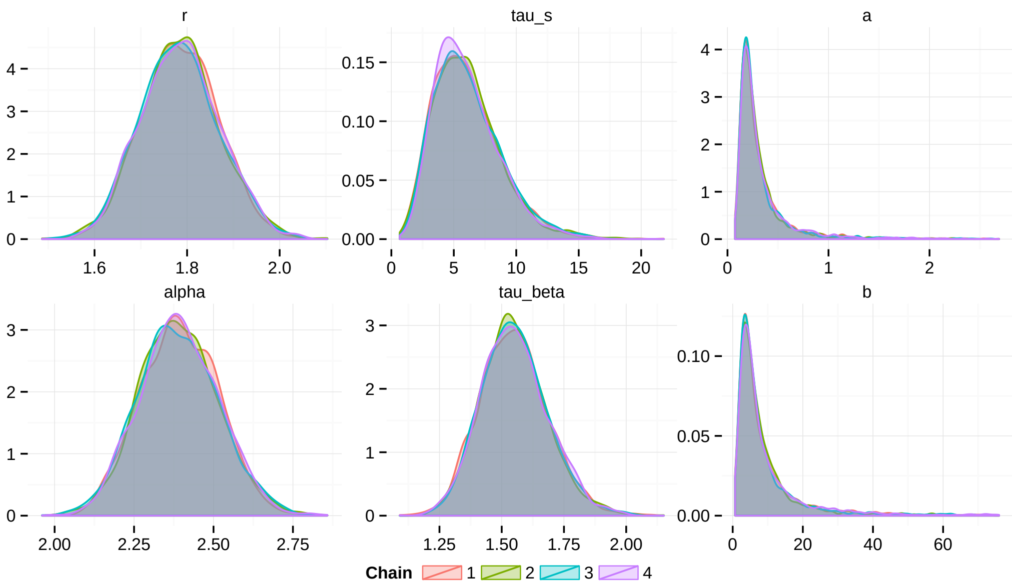

Figure 8.12: HSMDO/Apparel Marginal Posterior Densities for r, α, τ_s, τ_β, a, and b after 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

8.4.2 Results and Forecast Accuracy

8.4.2.1 Seasonal Components and Individual Seasonality

The HMC simulation with the no-U-turn sampler for the HSMDO resulted in seasonal and customer lifetime information on a cross-sectional as well as on an individual customer level. I obtained estimates for the overall seasonal components $s_1 - s_{12}$ and the individual seasonality coefficients β_i for all datasets (see Tables 8.1-8.3). The estimates for customers individual seasonality β_i and purchase rate λ_i are compiled in form of histograms in Figure 8.1 (DIY), Figure 8.3 (Apparel), and Figure 8.5 (CDNOW).

The seasonality measure β_i can be used for targeting customer groups and customer segmentation. Figure 8.13 shows an exemplary segmentation for the three retail customer bases into customers that show amplified seasonality $\beta_i \geq 1$, customers that show attenuated seasonality with β_i between 0 and 1, and anti-seasonal customers with $\beta_i < 0$. The comparison between tbe HSMDO segmentation to

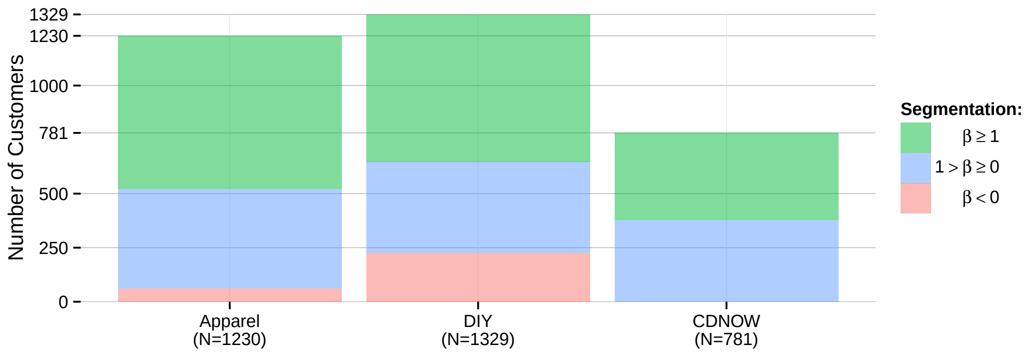

Figure 8.13: HSMDO Segmentation of Customer Base by Individual Seasonality β_i for each Retailer

the segmentation based on the HSM estimates (compare Figure 7.10, p. 111 and histograms in Figures 7.1-7.3, pp. 103-105) reveals that while the number of DIY customers in each segment remain almost identical, the possibility of drop-out induces a shift in estimated customer segments for the apparel and CDNOW customer base. The apparel retailers segment of pro-seasonal customers grows (HSM: 661 vs. HSMDO: 709), while the CDNOW estimates shows the opposite effect: the addition of the drop-out process leads to a smaller pro-seasonal customers segment (HSM: 531 vs. HSMDO: 403).

In the following I analyze the estimates for overall and individual seasonality separately for each retailer in more detail.

Estimates for the DIY Retailer

The overall estimated seasonal components $s_1 - s_{12}$ (Table 8.1) for the DIY Retailer indicate lower overall purchase levels from s_4 (November) to s_8 (March) and higher overall purchases level from s_9 (April) to s_3 (October). Compared to the HSM, the HSMDO provides nearly identical estimates for the overall seasonal components, showing the same seasonal pattern.

The customers' purchase frequencies are on average down approximately[31] 24%

[31] The seasonal components can be interpreted as the approximate (for small values) percentage change compared to the seasonal cycle average. This approximation is used throughout this thesis.

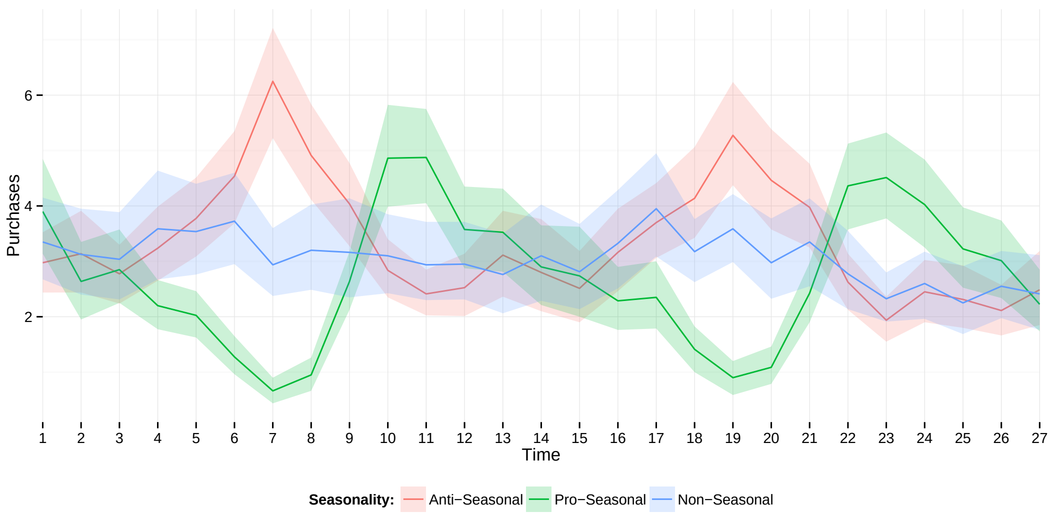

Figure 8.14: HSMDO/DIY Purchase History Comparison of Customer Groups Selected by Individual Seasonality Estimates β_i

in s_7 (February) and up 15% in s_{10} (May) from their yearly mean. This overall pattern indicates that the DIY retailer's customers activity peaks in the summer, maybe because customers use the summer holidays for home improvement or because the summer weather facilitates outdoor projects.

While overall seasonal information could have been obtained through other methods as well, the HSMDO also — and more importantly — yields measures of seasonality that reflect how each customer's individual behavior relates to the seasonal behavior of other customers. In contrast to the HSM's measures of seasonality the HSMDO's estimates of individual seasonality take potential customer drop-out into account. The discriminative power of this measure is illustrated in Figure 8.14. From the DIY retailer's customer base I selected three groups with 50 customers each: a pro-seasonal (green), an anti-seasonal (red) and a non-seasonal (blue) customer group. The pro-seasonal group comprises the customers with the largest β_i coefficients, the anti-seasonal group consists of customers with the lowest (negative) β_i coefficients, and for the non-seasonal group, customers with β_i closest to zero were selected. The shaded area around the lines depict the 95% confidence limits of the group mean computed by a nonparametric bootstrap.

The group graphs in Figure 8.14 show clearly that the HSMDO captures individual seasonality quite well. Over time, the pro-seasonal customers' actual purchase levels are an almost exact mirror image of the anti-seasonal customers levels, while the non-seasonal customers remain buying at fairly constant purchase levels. For example, the upward spikes in time periods 7 and 19 of the anti-seasonal customers is matched by the downward spikes of the pro-seasonal customers in the same time periods. This actual behavior corresponds with the estimated overall seasonal component s_7 (February), that indicates an overall reduction in purchase levels by 24% for periods 7 and 19. The pro-seasonal group's upward spike in time periods 11 and 23 is consistent with the +13% increase in purchase levels implied by component s_{11} (June).

Compared to the HSM, the HSMDO yields very similar results, however the HSMDO discriminates the groups slightly better. For example, the HSM confidence intervals for time period 19 in Figure 7.11 (p. 112) show some overlap between the non-seasonal and anti-seasonal customers. The HSMDO confidence intervals, in comparison, do not overlap so that the groups for that time period are fully discriminated. These difference are very slight.

Estimates for the Apparel Retailer

The seasonal components $s_1 - s_{12}$ (Table 8.2) for the apparel dataset indicate, that the highest overall purchase levels are to be expected in January (+101%; $s_1 = 0.6991$), December (+66%; $s_{12} = 0.5089$) and July (+35%; $s_7 = 0.2970$), which coincides with the summer and winter holiday seasons in the retailer's market. The lowest purchase levels are estimated in September (-52%; $s_9 = -0.7375$) and February (-41%; $s_2 = -0.5257$). Compared to the apparel estimates for the HSM (Table 7.2, p. 104) the overall seasonal components follow the same pattern but the differences between the models are larger than on the DIY data.

The HSMDO estimate for the precision parameter τ_β (1.55) of the apparel customers' seasonality is larger than the same estimate for the DIY retailer's customers (0.46). This pattern also emerges by comparing the spread of the histograms for β depicted in Figures 8.1 and 8.3. This result implies that the indi-

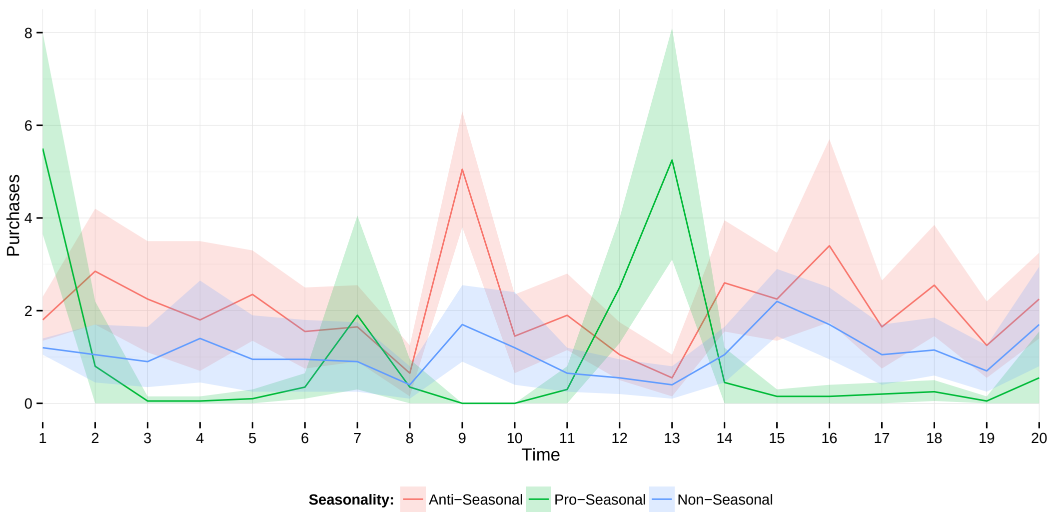

Figure 8.15: HSMDO/Apparel Purchase History Comparison of Customer Groups Selected by Individual Seasonality Estimates β_i

vidual β_i are more concentrated for the apparel retailer. One reason might be that apparel data spans a total time frame of 24 months and therefore, with reserving four months for hold-out, provides only eight month seasonal overlap for seasonal learning — less than the DIY data. Nonetheless, both the apparel and DIY estimates provide more than adequate discrimination of individual seasonality.

Figure 8.15 depicts a comparison of actual purchase histories of three customer groups selected by β_i estimates, employing the same procedure as above. Although not as clear as on the DIY retailer dataset, the pro- (green) and anti-seasonal (red) graphs mirror each other markedly over the time line. For example, the peak in pro-seasonal behavior at time periods 1 and 13 (s_1) coincides with the corresponding slump in the anti-seasonal group, equally the peak in anti-seasonal activity in period 9 (s_9) is mirrored by a simultaneous slump in purchase levels of the pro-seasonal group. The non-seasonal customers (blue) are somewhat in between with attenuated seasonal behavior. Comparing HSMDO group behavior to the graphs based on the HSM estimates (Figure 7.12, p. 113) suggests only minor differences, for example the HSMDO estimates' peak for the pro-seasonal customers in time period 7 is more subdued.

Estimates for CDNOW

The CDNOW dataset provides only little information for seasonal learning as it only spans 18 months in total. With four months of data reserved as hold-out that only leaves 14 months for learning and thus two months overlap for seasonal inference. It is noteworthy that under these conditions, the model still produces sensible individual estimates for all parameters (see Table 8.3). This is one of the strength of hierarchical models because the estimates in view of sparse data are pulled toward their higher level means. Thus, the individual estimates for β_i are highly concentrated ($\tau_\beta = 739$) around one (compare the histogram in Figure 8.5).

The lack of variance in the individual seasonality estimates β_i is reflected in the group comparison depicted in Figure 8.16. The model could only discern pro-seasonal (green) and least-seasonal (red) customers as no values for β_i were near zero or even negative. The least-seasonal graph only vaguely moves opposite to the pro-seasonal graph, with slumps coinciding with the corresponding peaks, for example, in time periods 6 and 13. However, this is inconclusive and judging from the high concentration of β_i around one, the majority of the seasonal learning for the CDNOW dataset will come from the overall seasonal components $s_1 - s_{12}$,

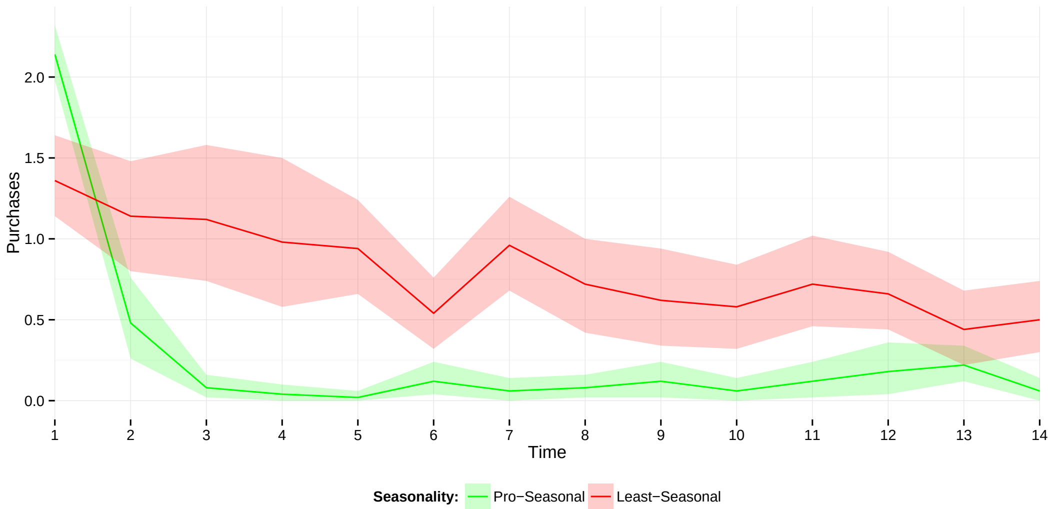

Figure 8.16: HSMDO/CDNOW Purchase History Comparison of Customer Groups Selected by Individual Seasonality Estimates β_i

with very little contribution by the individual seasonality estimates.

According to the estimates (see Table 8.3) for the seasonal components $s_1 - s_{12}$, peak activity for CDNOW is in November ($s_{11} = 0.1319$) through January ($s_1 = 0.9469$), while the lowest levels are to be expected in April ($s_4 = -0.2244$) and May ($s_5 = -0.2185$). The estimated overall seasonal components differ not only in value from the HSM (Table 7.3, p. 105), but also in sign. For example, while the HSM estimates higher than average sales in $s_1 - s_3$, the HSMDO estimates higher than average sales in $s_{11} - s_1$ and lower average sales in s_2 and s_3.

8.4.2.2 Prediction of Customer Inactivity

Predicting whether a customer is about to become inactive, or not, is a classification task. The ROC analysis shows that $P(Zero_F)$ dominates $P(alive)$ over all cutoff values and that $P(Zero_F)$ performs substantially better than the heuristic. This holds true for all three retailers and is corroborated by the superior AUC of $P(Zero_F)$ compared to the other methods. I will present the results in detail below.

I used Equation (8.25) during the HMC simulation of the HSMDO to calculate $P(alive)$ for each individual customer (see Section 8.2.2). I used Equation (8.41, p. 142) during the HMC simulation of the HSMDO to calculate $P(Zero_F)$ for each individual customer with F set to the length of the hold-out period (four months). The hiatus heuristic is described in Section 3.3.2.

Results for the DIY Retailer

The ROC curves in Figure 8.17 show that $P(Zero_F)$ (red) dominates $P(alive)$ (blue) over the complete range of cutoff values for the DIY retailer dataset. A curve drawn out further toward the upper left corner implies better classification results. Also, $P(Zero_F)$ quite substantially outperforms the hiatus heuristic (green) over a wide range of cutoff values. The area under the curve (AUC) for $P(Zero_F)$ is 0.8854 compared to 0.7871 for the hiatus heuristic (larger is better). $P(alive)$ is worst with an area under the curve of 0.7737.

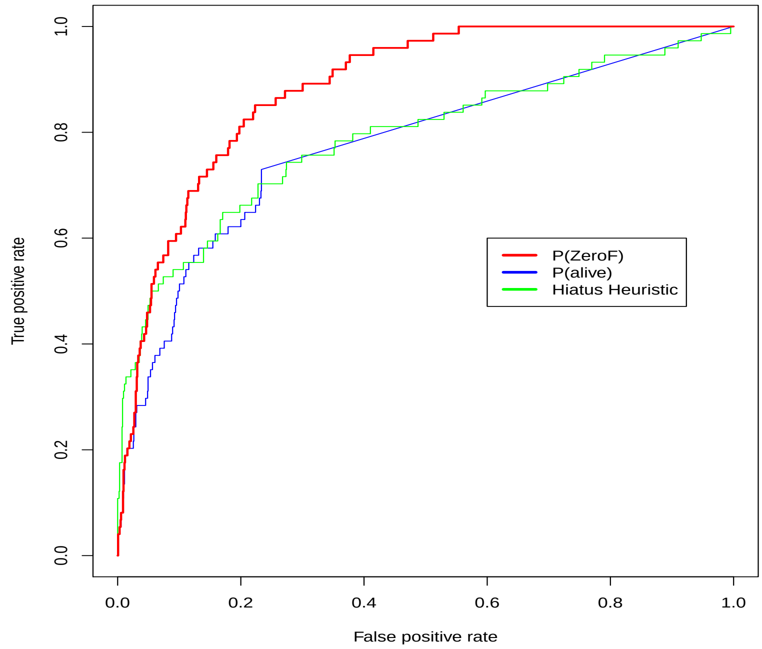

Figure 8.17: HSMDO/DIY ROC Analysis of Predicting Customer Inactivity with $P(Zero_F)$, $P(alive)$, and the Hiatus Heuristic

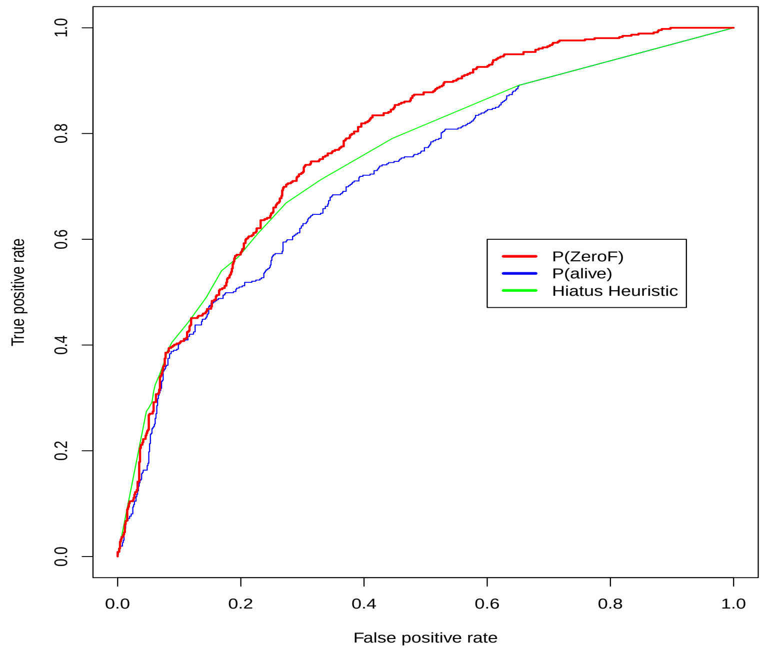

Figure 8.18: HSMDO/Apparel ROC Analysis of Predicting Customer Inactivity with $P(Zero_F)$, $P(alive)$, and the Hiatus Heuristic

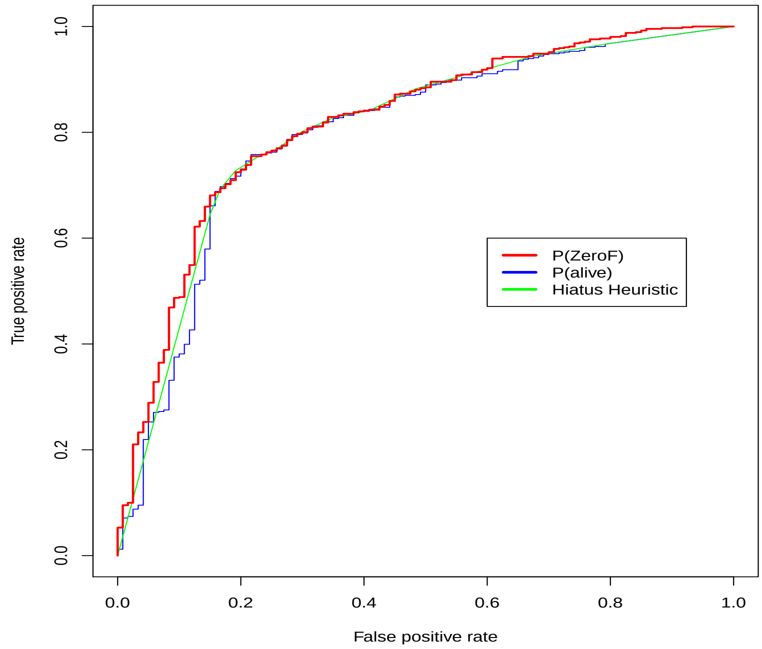

Figure 8.19: HSMDO/CDNOW ROC Analysis of Predicting Customer Inactivity with $P(Zero_F)$, $P(alive)$, and the Hiatus Heuristic

Results for the Apparel Retailer

The ROC analysis for the apparel retailer is depicted in Figure 8.18. Again, the ROC of $P(Zero_F)$ dominates the ROC of $P(alive)$ over the full parameter space and also substantially outperforms the hiatus heuristic over a wide range of cutoff values. This visual impression is corroborated by the corresponding area under the curve measures. Predictions based on $P(Zero_F)$ yield the largest AUC with 0.7809, surpassing the hiatus heuristic with an AUC of 0.7499 and $P(alive)$ with an AUC of 0.7202.

Results for the CDNOW Retailer

The classification results for CDNOW are much closer as can be seen in Figure 8.19. Still, the ROC curve for $P(Zero_F)$ dominates the ROC of $P(alive)$ over all cutoff values. While not completely dominating, $P(Zero_F)$ outperforms the hiatus heuristic over almost the full parameter space with only few exceptions. In line with the results for the other datasets, $P(alive)$ yielded the least accurate classification results. The area under the curve metric mirrors these results. $P(Zero_F)$

Table 8.4: ROC/AUC Analysis of Predicting Customer Inactivity

	DIY Retailer	Apparel Retailer	CDNOW
Hiatus Heuristic:	0.7871	0.7499	0.8053
HSMDO $P(alive)$:	0.7737	0.7202	0.7976
HSMDO $P(Zero_F)$:	0.8854	0.7809	0.8165

classification accuracy comes out ahead with 0.8165, followed by the heuristic with 0.8053 and $P(alive)$ in last place with 0.7976. The AUC measures for all datasets are compiled in Table 8.4.

8.4.2.3 Long- and Short-Term Predictions

The HSMDO generates the most accurate forecasts both long- and short-term. These results are consistent over all three retail datasets. The HSMDO and HMDO, the models with drop-out, yield substantially better forecasts for the CDNOW dataset than the other methods. The gap in forecast accuracy is closest on the apparel data, for which the HSM achieves nearly the same accuracy as the HSMDO. The HSMDO, HMDO and HSM are in general more accurate than the heuristic, SARIMA, and Holt-Winters methods on all retail data tested. These results are explained in detail below.

The following models are included in the comparison: the basic hierarchical seasonal model without drop-out (HSM), the full hierarchical seasonal model with drop-out (HSMDO), and a variant of the HSMDO without the seasonal structure but with drop-out (HMDO). The models are compared to Holt-Winters method, SARIMA model, and the baseline heuristic, that simply assumes that:

> *Customers continue to buy at their mean past purchase frequency.*

Results for the DIY Retailer

Table 8.5 compiles the long-term forecast accuracy results for the DIY retailer. The HSMDO, the model that includes both the seasonal structure and the drop-out process, outperforms all other methods and models with the lowest errors on all metrics. Interestingly, the HMDO produces a lower *mean MSE* (5.43) when

Table 8.5: DIY Retailer - Seasonal Model Long-Term Prediction Accuracy

	HW	SARIMA	Heuristic	HSM	HMDO	HSMDO
MAE/MAD	2.335	1.702	1.696	1.636	1.653	1.597
Mean MSE	11.751	5.948	5.607	5.518	5.430	5.376
Median MSE	4.550	2.520	2.536	2.324	2.433	2.251
Mean RMSE	2.717	1.957	1.940	1.894	1.900	1.857
Median RMSE	2.133	1.587	1.592	1.524	1.560	1.500

Table 8.6: DIY Retailer - Seasonal Model Short-Term Prediction Accuracy

	HW	SARIMA	Heuristic	HSM	HMDO	HSMDO
MAE/MAD	10.293	5.885	1.647	1.634	1.601	1.589
Mean MSE	112.750	39.451	5.182	5.107	4.979	4.954
Median MSE	114.993	36.894	1.494	1.491	1.398	1.354
Mean RMSE	10.293	5.885	1.647	1.634	1.601	1.589
Median RMSE	10.723	6.074	1.222	1.221	1.182	1.164

compared to the HSM (5.518), but a higher overall MAE/MAD (1.653 vs. 1.636). This implies that both the drop-out process and the seasonal structure contribute to the superior performance of the HSMDO. While not as good as the HSMDO, both the HMDO and HSM are more accurate than the heuristic, SARIMA model, and Holt-Winters method. The SARIMA model outperforms the heuristic on the less outlier sensitive median aggregated individual forecasts ($median\ MSE$: 2.520 vs. 2.536).

The short-term forecast accuracy for the DIY retailer is shown in Table 8.6 and confirms the results of the long-term forecasts. Again, the HSMDO is best on all metrics, producing the fewest errors. The HMDO is more accurate than the HSM, and in turn the HSM outperforms the heuristic. The worst short-term forecast accuracy resulted from the Holt-Winters and SARIMA methods. In general, the long-term forecasts show greater deviations than the short-term forecasts.

Results for the Apparel Retailer

The apparel retailers dataset spans 24 months of data which provides less seasonal overlap than the DIY retail data but more seasonal overlap than the CDNOW dataset. The forecast results for the long-term predictions are shown in Table 8.7. The HSMDO performs better than Holt-Winters (HW) method, SARIMA

Table 8.7: Apparel - Seasonal Model Long-Term Prediction Accuracy

	HW	SARIMA	Heuristic	HSM	HMDO	HSMDO
MAE/MAD	1.254	0.887	0.897	0.884	0.861	0.839
Mean MSE	4.488	2.732	2.514	2.509	2.634	2.510
Median MSE	1.374	0.563	0.610	0.536	0.466	0.466
Mean RMSE	1.570	1.104	1.105	1.104	1.091	1.090
Median RMSE	1.172	0.750	0.781	0.732	0.682	0.681

Table 8.8: Apparel - Seasonal Model Short-Term Prediction Accuracy

	HW	SARIMA	Heuristic	HSM	HMDO	HSMDO
MAE/MAD	1.449	1.708	0.874	0.786	0.844	0.764
Mean MSE	3.312	3.948	2.367	2.306	2.417	2.321
Median MSE	1.855	3.063	0.250	0.072	0.210	0.046
Mean RMSE	1.449	1.708	0.874	0.786	0.844	0.764
Median RMSE	1.362	1.750	0.500	0.269	0.458	0.215

models, the heuristic and the HMDO on all error metrics. The HSMDO is also superior to the HSM, except for the *mean MSE* (2.509 vs. 2.510), where both are approximately equal. The results between the HSM and HMDO are again mixed, with the HMDO providing more accurate overall forecasts according to MAE/MAD (0.861 vs. 0884) while the HSM is more accurate on the outlier sensitive *mean MSE* (2.509 vs. 2.634).

The short-term forecasts compiled in Table 8.8 provide much of the same picture. HSMDO performs better than Holt-Winters method, SARIMA models, the heuristic and the HMDO on all error metrics. The HSMDO is also superior to the HSM, except for the *mean MSE* (2.306 vs. 2.321). The HSM in turn superior is to all other methods including the HMDO.

Table 8.9: CDNOW - Seasonal Model Long-Term Prediction Accuracy

	HW	SARIMA	Heuristic	HSM	HMDO	HSMDO
MAE/MAD	NA	0.131	0.206	0.170	0.100	0.088
Mean MSE	NA	0.122	0.125	0.089	0.083	0.083
Median MSE	NA	0.000	0.005	0.008	0.000	0.000
Mean RMSE	NA	0.157	0.225	0.194	0.130	0.122
Median RMSE	NA	0.000	0.071	0.090	0.001	0.001

Table 8.10: CDNOW - Seasonal Model Short-Term Prediction Accuracy

	HW	SARIMA	Heuristic	HSM	HMDO	HSMDO
MAE/MAD	NA	0.347	0.222	0.223	0.152	0.135
Mean MSE	NA	0.197	0.157	0.146	0.140	0.141
Median MSE	NA	0.082	0.005	0.013	0.000	0.000
Mean RMSE	NA	0.347	0.222	0.223	0.152	0.135
Median RMSE	NA	0.286	0.071	0.113	0.001	0.001

Results for CDNOW

The CDNOW dataset only includes 18 months of total data. This, with four months of data reserved as hold-out, leaves only 14 months for parameter learning, implying as few as two month overlap for seasonal inference. Holt-Winters method did not yield any results as the overlap was too small — this is denoted as *NA*. The long- and short-term forecast results for the CDNOW dataset are compiled in Tables 8.9 and 8.10.

The drop-out process seems to capture an essential part of the behavior of the CDNOW customer base as in general both the HSMDO and HMDO provide substantially better forecasts compared to the other methods including the HSM. For example, according to long-term MAE/MAD the models rank as follows: 0.088 (HSMDO), 0.100 (HMDO), 0.131 (SARIMA), 0.170 (HSM), 0.206 (heuristic). This pattern holds for both the short- and long-term. Further study of the CDNOW dataset reveals that it contains – compared with the other datasets – a high number of customers with their last purchase dating back a relatively long time. This makes the forecasts for this dataset very sensitive to customer drop-out and trend components.

It is noteworthy that the SARIMA model for long-term forecasts performs better than the heuristic on all error metrics. Here, the SARIMA model produces a *median MSE* and *median RMSE* of zero on par with the HMDO and HSMDO[32] An explanation for this is that the SARIMA procedure identifies downward trends for a number of customers, emulating a form of drop-out. The HSMDO provides equal or better long-term forecasts than the HMDO. This holds also true in the

[32] The difference in *median RMSE* between 0.000 and 0.001 is negligible.

short-term with the only exception of *mean MSE* that is negligibly lower for the HMDO (0.140 vs. 0.141).

8.4.2.4 Prediction of Future Best Customers

The results of forecasting the future top customers show that the HSMDO is equal or better than the other methods for the DIY and CDNOW data, while the HSM is best for the apparel data. However, the differences are small, so that none of the methods set itself apart. The findings are explained in detail below.

Rank ordering provides a different view on forecast accuracy compared to the error metrics in the previous section. For example, even if forecast errors are quite large for a particular method, it could still be that customer purchase levels are captured accurately relative to other customer purchase levels. This is an important property for many essential segmentation and customer prioritization tasks a marketing manager might perform.

As in the previous section I benchmark the HSMDO against the heuristic, the HSM, and the HMDO. The heuristic simply assumes that:

> *The past 10% (20%) best customers will also be the future 10% (20%) best customers.*

Results for the DIY Retailer

The results for the DIY retailer for predicting the top 10% and 20% of the customer base are presented in Table 8.11. Interestingly the HMDO and HSMDO

Table 8.11: DIY Retailer - Accuracy of Top 10% (20%) Customer Prediction

(%)	**Heuristic**		**HSM**		**HMDO**		**HSMDO**	
True Positives:	54.55	(61.89)	54.55	(62.26)	56.82	(63.39)	56.82	(64.53)
True Negatives:	94.99	(90.51)	94.99	(90.60)	95.24	(90.88)	95.24	(91.17)
False Negatives:	45.45	(38.11)	45.45	(37.74)	43.18	(36.61)	43.18	(35.47)
False Positives:	5.01	(9.49)	5.01	(09.40)	4.76	(9.12)	4.76	(8.83)
Error:	9.03	(15.20)	9.03	(15.05)	8.58	(14.60)	8.58	(14.15)
Accuracy:	90.97	(84.80)	90.97	(84.95)	91.42	(85.40)	91.42	(85.85)

Table 8.12: CDNOW - Accuracy of Top 10% (20%) Customer Prediction

(%)	Heuristic		HSM		HMDO		HSMDO	
True Positives:	41.03	(42.31)	41.03	(42.95)	42.31	(46.15)	42.31	(46.15)
True Negatives:	93.46	(85.60)	93.46	(85.76)	93.60	(86.56)	93.60	(86.56)
False Negatives:	58.97	(57.69)	58.97	(57.05)	57.69	(53.85)	57.69	(53.85)
False Positives:	6.54	(14.40)	6.54	(14.24)	6.40	(13.44)	6.40	(13.44)
Error:	11.78	(23.05)	11.78	(22.79)	11.52	(21.51)	11.52	(21.51)
Accuracy:	88.22	(76.95)	88.22	(77.20)	88.48	(78.49)	88.48	(78.49)

Table 8.13: Apparel - Accuracy of Top 10% (20%) Customer Prediction

(%)	Heuristic		HSM		HMDO		HSMDO	
True Positives:	52.85	(59.76)	53.66	(60.16)	51.22	(57.72)	52.85	(59.76)
True Negatives:	94.76	(89.94)	94.85	(90.04)	94.58	(89.43)	94.76	(89.94)
False Negatives:	47.15	(40.24)	46.34	(39.84)	48.78	(42.28)	47.15	(40.24)
False Positives:	5.24	(10.06)	5.15	(09.96)	5.42	(10.57)	5.24	(10.06)
Error:	9.43	(16.10)	9.27	(15.93)	9.76	(16.91)	9.43	(16.10)
Accuracy:	90.57	(83.90)	90.73	(84.07)	90.24	(83.09)	90.57	(83.90)

yield identical performances in identifying the future top 10% of customers, both being 91.42% accurate. This is more accurate than the heuristic and HSM, which, again, yield identical results (90.97%) for the top 10% of the customer base. One explanation for this result is, that the absolute top customers seem to purchase at relatively high constant rates, exhibiting low seasonality.

Regarding the prediction of the top 20%, the seasonal HSMDO is more accurate than the HMDO (85.85% vs. 85.40%) and also outperforms the other methods. The HSM is slightly ahead of the heuristic (84.95% vs. 84.80%). The seasonal components seem to yield more forecast relevant information for the top 20% than for the top 10% of customers. The SARIMA model and Holt-Winters method yield inferior results (see Table 7.10, p. 120).

Results for CDNOW

For the CDNOW retailer data the HSMDO and HMDO yielded identical accuracy for both the top 10% (88.48%) and top 20% (78.49%) forecasts, ahead of HSM and the heuristic (see Table 8.12). The HSMDO and HMDO are also more accurate than the SARIMA model in predicting the top 10% as shown in Table

7.12, p. 120. The HSM edges out the heuristic in predicting the top 20% (77.20% vs. 76.95%), but yields identical results for the top 10% (88.22%). This supports the view that the CDNOW retailer forecasts benefit relatively more from the inclusion of a customer lifetime model and relatively less from the inclusion of seasonal effects. This is not surprising as the dataset features the least seasonal overlap.

Results for the Apparel Retailer

The model results for the apparel retailer (see Table 8.13) show that the HSM performs best, even better than the models that include drop-out. This holds for both the top 10% (90.73% accuracy) and top 20% (84.07% accuracy) forecasts. The HMDO generates the least accurate forecasts, while the HSMDO and heuristic provide identical results for the top 10% (90.57%) and top 20% (83.90%).

8.5 Discussion of Results

The HSMDO operates under the "buy 'til you die" assumption and contributes to the literature stream of probability models for customer base analysis (Fader and Hardie 2009). It is unique in that it combines (1) an explicit customer lifetime model, (2) individual and cross-sectional drop-out rates, (3) individual purchase rates and cross-sectional heterogeneity, (4) multiplicative seasonal effects, (5) individual customer's seasonal behavior, and (6) a measure of customer inactivity in finite time horizons $P(Zero_F)$.

Similar to the PDO/NBD model (Jerath, Fader, and Hardie 2011), I assume discrete periodic death opportunities tied to calendar time. My model differs in that it is based on a discretely sampled inhomogeneous Poisson counting process, so that death opportunities coincide with the granularity of the purchase frequency sampling. Hybrid/Hamilton Monte Carlo Sampling (Brooks et al. 2011, pp. 113-160) in form of the no-U-turn sampler (Homan and Gelman 2014) is used to estimate model parameters instead of Gibbs sampling that was used for the DMPT and HSM. Hamilton Monte Carlo sampling substantially increases the effective sample size obtained in the same amount of computational time.

The HSMDO, like the HSM, includes individual and cross-sectional seasonal effects. It generalizes post-hoc seasonal adjustment (Zitzlsperger, Robbert, and Roth 2009) and dummy-variable approaches (Schweidel and Knox 2013) to improve forecast accuracy. The analysis of the HSMDO shows that it captures both individual and cross-sectional seasonality quite well — see Section 8.4.2.1. The results of targeting customer groups (as shown for example in Figures 8.14, p. 156) and segmenting the customer base (see Figure 8.13, p. 8.13) according to the model's seasonal estimates demonstrate that this measure has a high discriminative power, even with datasets that feature only moderate seasonal overlap.

For the prediction of inactive customers, I derived a new metric $P(Zero_F)$ for the HSMDO that outperforms the commonly used $P(alive)$ metric first derived by Schmittlein, Morrison, and Colombo (1987) for the Pareto/NBD. This metric improves upon $P(alive)$ by providing additional flexibility through the parameter F that specifies the length of the planning horizon, yielding a theoretical sound measure for finite time horizons, and providing superior predictive accuracy (see Section 8.3, p. 136). The ROC analysis in Section 8.4.2.2 demonstrates that $P(Zero_F)$ completely dominates $P(alive)$ over all cutoff values in all three retail settings. $P(Zero_F)$ also outperforms the hiatus heuristic over a wide range of cutoff values, yielding by far the largest AUC on all three datasets.

Not only does $P(Zero_F)$ provide more accurate forecasts it also provides distinct information about managerial relevant outcomes. The reason for this is that $P(alive)$ pertains to an infinite time horizon, but in most managerial decision situations the forecast horizon ranges from the next quarter, the next year to maybe the next five years at most. However, this does not imply that $P(Zero_F)$ should fully replace the $P(alive)$ metric. $P(Zero_F)$ should be seen as augmenting $P(alive)$ for the HSMDO in situations where empirical validation with finite hold-out periods is performed or when the managerial context requires a certain planning horizon.

I compared the long- and short-term forecast accuracy of the HSMDO, the HSM, a simple heuristic, SARIMA models, and Holt-Winters method. I also included the HMDO, which is a variant of the HSMDO without the seasonal submodel, to gain further insights into which part of the model has the largest effect on predictive performance. The HSMDO generates the most accurate forecasts both in the long- and the short-term. These results were consistent over all retail data. Only on the apparel data, the HSM provides results comparable to the models with drop-out. The results of the HMDO, while not as good as the HSMDO, imply that both the drop-out process and the seasonal structure contribute to the superior performance of the HSMDO. Also, as expected, the seasonal models' performances improve with increased seasonal overlap in the data — as with the DIY and apparel retail data. The CDNOW data features only little seasonal overlap for learning and comprises a large number of customers who did not purchase recently. Here, the HSMDO and HMDO, the models with drop-out yield substantially better forecasts than all other methods. In general, the HSMDO and HSM achieve a higher short- and long-term forecast accuracy than the heuristic, SARIMA model, and Holt-Winters method for all three retailers.

The ability to identify the future top 10% or 20% of the customer base is another essential classification task. The results of the empirical validation suggest that even though the HSMDO was equal or better than the other methods for the DIY and CDNOW data, and the HSM was best for the apparel data, the differences are so slight that none of the methods set itself apart. These results mirror the findings of Wübben and von Wangenheim (2008) in a benchmark study as well as the simulation results by Huang (2012) for the Pareto/NBD. A reason for the HSMDO not setting itself further ahead of the other methods in this classification task might be that drop-out rates of actual top customers are distinctly lower than those of customers with average purchase frequencies. This implies that actual drop-out and purchase rates are not independent and that higher individual purchase rates are associated with lower drop-out rates than the model currently predicts. Thus, one could hypothesize that "best customers don't die easily".

Chapter 9

Summary and Conclusions

9.1 Summary of Central Results

This thesis contributes to contemporary research on probability models for customer base analysis (Fader and Hardie 2009). I modified one already existing model and developed two new models to further the understanding of the interrelation between individual and cross-sectional purchase rates, individual and cross-sectional seasonality, individual and cross-sectional drop-out rates, customer inactivity, and forecast accuracy. Thus, I proposed a hierarchical Bayesian seasonal effects model (HSM), a hierarchical Bayesian seasonal effects model with drop-out (HSMDO), and a refined rule to predict customer inactivity for the dynamic model of purchase timing (DMPT). In addition, I derived a measure $P(Zero_F)$ for the HSMDO that is superior to the $P(alive)$ metric commonly used in marketing models with an underlying customer lifetime model. In detail, I made the following contributions:

I proposed a new refined rule to predict customer inactivity for the DMPT that takes into account the actual predictive distribution and finite length of the hold-out period. It yields a slight but consistent improvement on previous approaches. The ROC analysis shows that the new rule is superior over all potential cutoff values on three different retail datasets. However, in comparison with managerial heuristics the results of the DMPT's performance in the simulated managerial tasks are mixed. In general, the DMPT performs relatively best in predicting long- and short-term interpurchase times while the heuristic is more accurate in predicting future top customers.

I newly developed the HSM that hierarchically combines multiplicative seasonal effects, individual and group-level seasonal behavior, and a measure of individual customers' seasonality. The model operates under the "always a share" assumption and extends the NBD model structure that relates individual purchase rates to cross-sectional heterogeneity (e.g., Schmittlein, Morrison, and Colombo 1987; Jerath, Fader, and Hardie 2011; Bemmaor and Glady 2012; Mzoughia and Limam 2014). The HSM generalizes and improves on post-hoc seasonal adjustment (Zitzlsperger, Robbert, and Roth 2009) and seasonal dummy-variable approaches (Schweidel and Knox 2013). My analysis shows that the model yields a measure of individual seasonality that has a high discriminative power, even with sparse datasets that feature only moderate seasonal overlap. This measure allows for targeting groups of customers according to their seasonality and segmenting the customer base into pro-seasonal, non-seasonal, and anti-seasonal customers. In comparison to simple heuristics, SARIMA models, and the Holt-Winters method, the HSM's forecast accuracy in simulated managerial tasks is superior for both the DIY and the apparel retailer dataset. For CDNOW's customer base the forecast results are not as clear as the HSM does not set itself apart on all error metrics. This is in part due to the minimal seasonal overlap the CDNOW data provides for seasonal learning, but also because the CDNOW data exhibits a large number of customers, who potentially have already dropped out.

I newly developed the HSMDO that combines an explicit customer lifetime model with multiplicative seasonal effects and relates individual to cross-sectional heterogeneity in drop-out rates, purchase rates, and seasonality. The drop-out process I propose is closely related to previous work on the Pareto/NBD model family under the "buy 'til you die" assumption (e.g., Schmittlein, Morrison, and Colombo 1987; Jerath, Fader, and Hardie 2011; Bemmaor and Glady 2012; Mzoughia and Limam 2014; Glady, Lemmens, and Croux 2015). I assume discrete periodic death opportunities tied to calendar time based on a discretely sampled inhomogeneous Poisson counting process. In this framework drop-out opportunities coincide with the granularity of the purchase frequency sampling. HSMDO parameter estimates and predictions were obtained through Hybrid/Hamilton Monte Carlo Methods (Brooks et al. 2011, pp. 113-160) in form of the no-U-turn sampler (Homan and Gelman 2014). The analysis of the HSMDO shows that it yields a measure of individual seasonality that has a high discriminative power (even higher than the HSM) for datasets that feature moderate seasonal overlap. A comparison of the HSMDO, the HSM, the HMDO, a simple heuristic, SARIMA models, and the Holt-Winters method demonstrates that the HSMDO yields the most accurate forecasts both in the long- and short-term for all three retail datasets. The HSMDO and HMDO, the models with drop-out processes, yield substantially better forecasts for CDNOW's customer base compared to the models without an explicit customer lifetime model. The seasonal models' (HSM and HSMDO) accuracy increases relative to the non-seasonal models with larger amounts of seasonal overlap present in the data.

I derived a new metric $P(Zero_F)$ to predict customer inactivity for the HSMDO. It is an alternative to the commonly used $P(alive)$ metric, first proposed by Schmittlein, Morrison, and Colombo (1987) for the Pareto/NBD model. $P(Zero_F)$ improves upon $P(alive)$ by (a) providing additional flexibility through the parameter F that specifies the length of the planning horizon, (b) yielding a theoretical sound measure for finite time horizons, and by (c) providing superior predictive accuracy. The ROC analysis demonstrates that $P(Zero_F)$ completely dominates $P(alive)$ over the complete space of potential cutoff values on all three retail datasets. It outperforms the hiatus heuristic by providing distinctly higher AUC measures for all three retailers. Moreover, $P(Zero_F)$ solves the conundrum that predictions based on $P(alive)$ cannot be empirically validated because $P(alive)$ pertains to an infinite time horizon.

9.2 Managerial Implications

With this thesis I strive not only for knowledge of academic interest, but also provide guidance for practitioners, who want to manage their customer base effectively in noncontractual settings. I focus on obtaining practically relevant information about individual customers and providing forecasts that facilitate the decision-making process in customer relationship management. This thesis addresses key managerial tasks for which marketing executives currently use probabilistic models or heuristics. Customer relationship managers have to be aware of the strengths and weaknesses of these methods, because their applicability and accuracy highly depend on the decision context. My research provides a benchmark that shows that the same model varies in its predictive performance over a number of real-life decision situations that, at first glance, seem strongly related.

I want to emphasize that managers should not only assess how well a model predicts the future, but also be aware of the kind of information it yields about the customer base. In the following, I first delineate specific managerial implications in regard to the seasonal information the models provide and then detail the potential implications of the forecast results in the three simulated managerial tasks analyzed in this thesis: the prediction of future customer inactivity, the forecast of purchase levels, and the prediction of future best customers.

Managerial Perspective on the Models' Seasonal Information

The new seasonal models I propose in this thesis, the HSM and HSMDO, yield seasonal information on both an overall customer base and an individual level. For example, the analysis of the DIY retailer's customer base indicates an overall seasonal pattern with customers' activity peaking in the summer. A reason for this pattern might be that customers use the summer holidays for home improvement or because the summer weather facilitates outdoor projects. The model estimates for the apparel retailer show that the overall highest purchase levels are expected in December, January, and July. These months coincide with winter holiday shopping season and summer holidays (in the retailer's market). While this information could have been easily obtained through other methods as well, the models also —

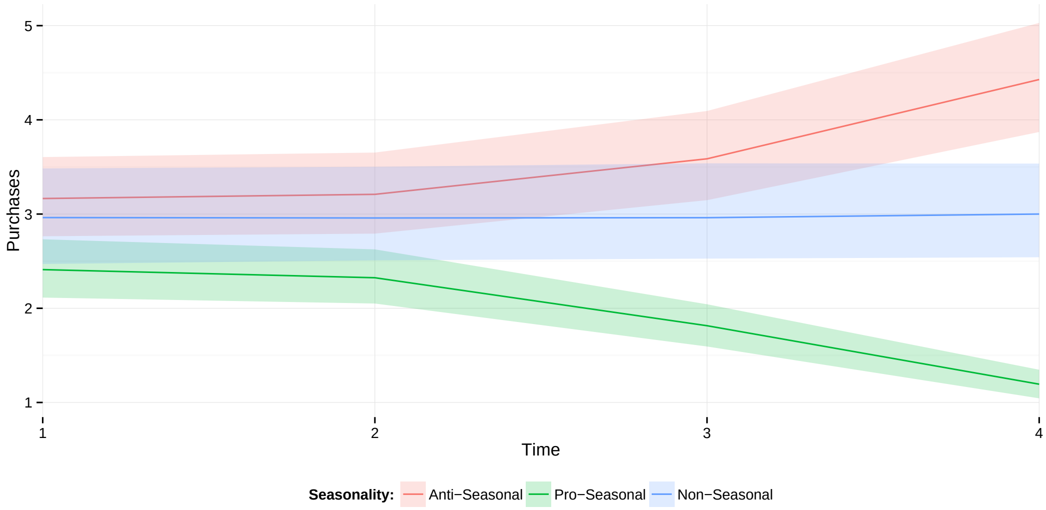

Figure 9.1: HSMDO/DIY Four Months Forecast of Average Sales Trajectories Based on Grouping Customers by Seasonality

and more importantly — yield an individual measure of seasonality that reflects how each customer's behavior relates to overall seasonal behavior.

This individual level seasonal information offers important insights for managers. For example, Figure 9.1 illustrates the differential seasonal targeting of three customer groups of the DIY retailer's customer base and the forecast of their purchase levels. The three groups are targeted based on how their seasonal purchase behavior relates to the overall seasonal pattern. The pro-seasonal group (green) comprises customers with strongest seasonality, the anti-seasonal group (red) consists of customers with negative seasonality. For the non-seasonal group (blue) I selected customers with a seasonality closest to zero.

Knowledge about an individual customer's seasonal purchase behavior can be used by marketing managers to improve the timing and effectiveness of targeted marketing actions. For example, cross-selling initiatives can be synchronized with peaks in customer activity. Selecting customers by seasonality could raise the success rate of retention programs by targeting only those customers that change their seasonal pattern.

Moreover, marketing actions that target non-seasonal customers could be stretched out more evenly across the year, allowing for more efficient marketing resource planning. Customer loyalty programs could reward pro- and anti-seasonal customers, who buy during their low activity periods. Executives might use this information to improve customer portfolio management. To mitigate risk they might even out seasonal peaks and slumps by attracting customers, who fit the desired seasonal profile.

Prediction of Future Customer Inactivity

Forecasts of customer inactivity are useful for the planning of marketing actions and the estimation of customer lifetime value. For example, proactive retention campaigns benefit from targeting individual customers, who will most likely make no purchases in the near future or in a certain planning horizon. If the strategic goal is not to retain potential defectors, but to increase cross-selling, a marketing manager might select those customers, who are most likely to purchase at least once in a certain future time frame.

The methods compared in this thesis differ vastly in their ability to accurately predict customer inactivity. The empirical validation clearly demonstrates the superiority of the HSMDO using the $P(Zero_F)$ metric over the hiatus heuristic and the commonly used $P(alive)$ metric on all three retail customer bases. Also, the analysis of the DMPT, a non-seasonal model without an explicit customer lifetime model, shows that such models may not yield any tangible advantage in predicting inactive customers, compared to a hiatus heuristic.

Marketing managers should be aware that $P(alive)$ pertains to an infinite time horizon, while in most managerial decision situations the forecast horizon ranges from the next quarter to the next year or maybe to the next five years at most. Thus, it seems odd that managers would use a such a measure when they are interested in the foreseeable future or a certain planning horizon. The alternative metric, $P(Zero_F)$, provides additional flexibility for practitioners through the parameter F that specifies the length of the planning horizon and superior predictive accuracy compared to $P(alive)$.

Forecast Accuracy of Purchase Levels

The ability to make individual level predictions of purchase behavior is essential for computing the future value of a firm's customers on a systematic basis and for assessing the potential return of investment of marketing campaigns (Fader and Hardie 2009).

The empirical analysis shows that stochastic models outperform the heuristic both for short- and long-term forecasts of purchase levels in all three retail scenarios. Practitioners should refrain from using SARIMA and Holt-Winters methods for individual customer forecasts with only sparse purchase histories as these methods performed consistently sub-standard.

The HSMDO generates the most accurate forecasts both in the long- and the short-term. The results are consistent over all retail datasets. Still, managers should carefully analyze the customer base and determine if their customers behave according to an "always a share" or to a "buy 'til you die" assumption. For example, the models with drop-out (HSMDO and HMDO) that operate under "buy 'til you die" paradigm yield substantially better forecasts for the CDNOW dataset, which has a high number of customers who did not purchase recently.

Prediction of Best Customers

Companies often rank-order their customers according to purchase frequencies to prioritize customer investments (Roberts and Berger 1999). The firm's top customers would normally be selected first for retention programs (Winer 2001; Shugan 2005), while customers that are believed to be unprofitable in the future might be abandoned (Haenlein, Kaplan, and Schoder 2006).

The results of the empirical validation of the predictive power suggest that even though the HSMDO was equal or better than the other methods for the DIY and CDNOW data and the HSM was best for the apparel data, the differences in accuracy are so slight that none of the methods sets itself apart from the heuristic. Thus, if marketing executives aim to identify future top customers, they should weigh the simplicity and easy-to-communicate nature of fast and frugal heuristics against a potential gain of more complex models which sometimes is limited.

9.3 Limitations and Implications for Future Research

As it is the case for virtually all research efforts, this work is subject to several limitations that need to be taken into account when evaluating the results. Furthermore, some findings raise new interesting questions that are outside the scope of this thesis, but should be addressed by future research.

The HSM and HSMDO have been validated for the datasets of an apparel retailer, a DIY retailer and CDNOW, an online retailer. These specific empirical settings might raise concerns about the generalizability of the inferred results. While I believe that the results extend to related noncontractual settings (e.g., grocery, books, furniture retailing) and hold for different product categories (e.g., food, household products, consumer electronics), I encourage researchers and practitioners to cross-validate the models in different retail settings.

Time periods covered by the data and the level of data aggregation used for parameter estimation may pose a limitation to model validity. For instance, additional validation might examine how the models cope with longitudinal data that spans more than the 31 months of DIY retail data used here. Also, I aggregated all customer data on a monthly purchase frequency level because the DIY retail data did not allow for finer granularity. Researchers might explore how finer or coarser granularity of data affects the forecast results of the models.

For model validation I chose the number of seasonal components to match the granularity of purchase frequency data (12 seasonal components). If data permits, one could use a larger number of seasonal components, e.g., 52 weekly components $s_1 - s_{52}$, or fewer components, e.g., quarterly components $s_1 - s_4$. I derived the model in full generality so that the number of seasonal components can be adjusted through the link function. It would be interesting to investigate the effect of the number of components on the models' forecast accuracy and on the discriminative power of the individual seasonality measure. Moreover, it is possible to view the number of components as an intrinsic model parameter that is drawn from a prior distribution (Xiao, Kottas, and Sansó 2015). Thus, further research might explore whether the HSMDO can be extended so that the optimal number of components is estimated from the data.

The HSMDO operates under a "buy 'til you die" assumption and the HSM under an "always a share" assumption. Recently, hidden Markov models have enabled researchers to relax these assumptions. Literature proposes non-seasonal models that fall in between these two assumptions and allow customers to switch between states of inactivity and activity (Mark et al. 2013; Romero, Van der Lans, and Wierenga 2013). It would be interesting to explore how the seasonal models presented in this thesis would benefit from a state-switching hidden Markov model.

Another fruitful research avenue pertains to the seasonal structure itself. For example, one could examine if the normal assumption of the seasonal components can be replaced or augmented by an autoregressive or moving average model of interrelated seasonal components or by a smoothing kernel (Dutta 2015). The hierarchical Bayesian structure of the HSMDO and HSM provides researchers with a framework to parsimoniously examine alternative submodels and estimate their parameters.

The metric $P(Zero_F)$ is a theoretically sound approach to capture the probability of customer inactivity in finite time horizons and proved superior to $P(alive)$ for the focal model of this thesis. I encourage researchers to compare the out-of-sample accuracy of $P(Zero_F)$ to that of $P(alive)$ in other model frameworks, empirical settings, and for different lengths of hold-out periods.

One of the patterns that emerged from the models' results was that "better" customers seem to be relatively less likely to drop-out. This pattern might imply that the drop-out and purchase rates are not independent and are influenced by other factors. In the realm of non-seasonal models, there has been some previous work in this regard. For example, Glady, Lemmens, and Croux (2015) use copulas to model the relationship between drop-out and transaction rates. Schweidel, Park, and Jamal (2014) propose a model that links non-purchase activities to drop-out rates. Such approaches could be extended to the HSMDO to capture the relationship between seasonality, drop-out, and purchase rates.

The new models I propose in this thesis, the HSMDO and HSM, may serve as a canvas on which other researchers can envision and explore the exciting relationships between purchase rates, drop-out rates, customer inactivity, and seasonality. They can use these models as a basis to build new models that incorporate additional aspects of purchase behavior. For example, the models can easily be augmented by time-varying and time-invariant co-variates, e.g., geographic, psychographic and demographic variables, customer specific attributes, or non-purchase activities. The models allow to analyze the effects of co-variates adjusted by individual and group level seasonal behavior. Thus, future research might identify effects that were previously hard to isolate and discriminate against a background of complex seasonal patterns.

Bibliography

Abe, Makoto (2009), "'Counting Your Customers' One by One: A Hierarchical Bayes Extension to the Pareto/NBD Model," *Marketing Science*, 28 (3), 541–553.

Allenby, Greg M., David G. Bakken, and Peter E. Rossi (2004), "The HB Revolution," *Marketing Research*, 16 (2), 20–25.

Allenby, Greg M., Robert P. Leone, and Lichung Jen (1999), "A Dynamic Model of Purchase Timing with Application to Direct Marketing," *Journal of the American Statistical Association*, 94 (446), 365–374.

Alon, Ilan, Min Qi, and Robert J. Sadowski (2001), "Forecasting Aggregate Retail Sales: A Comparison of Artificial Neural Networks and Traditional Methods," *Journal of Retailing and Consumer Services*, 8 (3), 147–156.

Anderson, Eugene W. and Vikas Mittal (2000), "Strengthening the Satisfaction-Profit Chain," *Journal of Service Research*, 3 (2), 107–120.

Andrews, Rick L., Imran S. Currim, Peter S.H. Leeflang, and Jooseop Lim (2008), "Estimating the SCAN*PRO Model of Store Sales: HB, FM or Just OLS?" *International Journal of Research in Marketing*, 25 (1), 22–33.

Aravindakshan, Ashwin, Olivier Rubel, and Oliver Rutz (2015), "Managing Blood Donations with Marketing," *Marketing Science*, 34 (2), 269–280.

Assael, Henry (1988), *Consumer Behavior and Marketing Action*, 3rd edition, Melbourne: PWS-Kent and Nelson.

Babin, Barry J. and Eric G. Harris (2012), *CB3*, Mason, OH: South-Western Cengage.

Ballings, Michel and Dirk Van den Poel (2015), "CRM in Social Media: Predicting Increases in Facebook Usage Frequency," *European Journal of Operational Research*, 244 (1), 248–260.

Batislam, Emine Persentili, Meltem Denizel, and Alpay Filiztekin (2007), "Empirical Validation and Comparison of Models for Customer Base Analysis," *International Journal of Research in Marketing*, 24 (3), 201–209.

Bauer, Hans H., Nicola E. Sauer, and Christine Becker (2006), "Investigating the Relationship Between Product Involvement and Consumer Decision-Making Styles," *Journal of Consumer Behaviour*, 5 (4), 342–354.

Bemmaor, Albert C. and Nicolas Glady (2012), "Modeling Purchasing Behavior with Sudden 'Death': A Flexible Customer Lifetime Model," *Management Science*, 58 (5), 1012–1021.

Berger, James O. and Robert L. Wolpert (1988), *The Likelihood Principle*, 2nd edition, Hayward, CA: Institute for Mathematical Statistics.

Bernhardt, Kenneth L., Naveen Donthu, and Pamela A. Kennett (2000), "A Longitudinal Analysis of Satisfaction and Profitability," *Journal of Business Research*, 47 (2), 161–171.

Berry, Michael J. and Gordon S. Linoff (2004), *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, New York, NY: Wiley.

Blattberg, Robert C. and John Deighton (1996), "Manage Marketing by the Customer Equity Test," *Harvard Business Review*, 74 (July/August), 136–144.

Blattberg, Robert C., Byung-Do Kim, and Scott A. Neslin (2008), *Why Database Marketing?*, New York, NY: Springer.

Blattberg, Robert C., Jacquely S. Thomas, and Gary Getz (2001), *Customer Equity: Building and Managing Relationships as Valuable Assets*, Cambridge, MA: Harvard Business School Press.

Boatwright, Peter, Sharad Borle, and Joseph B. Kadane (2003), "A Model of the Joint Distribution of Purchase Quantity and Timing," *Journal of the American Statistical Association*, 98 (463), 564–572.

Böckenholt, Ulf (1998), "Mixed INAR(1) Poisson Regression Models: Analyzing Heterogeneity and Serial Dependencies in Longitudinal Count Data," *Journal of Econometrics*, 89 (1–2), 317–338.

Bolton, Ruth N., Katherine N. Lemon, and Matthew D. Bramlett (2006), "The Effect of Service Experiences Over Time on a Supplier's Retention of Business Customers," *Management Science*, 52 (12), 1811–1823.

Bolton, Ruth N., Katherine N. Lemon, and Peter C. Verhoef (2004), "The Theoretical Underpinnings of Customer Asset Management: A Framework and Propositions for Future Research," *Journal of the Academy of Marketing Science*, 32 (3), 271–292.

Bonfrer, Andre, George Knox, Jehoshua Eliashberg, and Jeongwen Chiang (2007), *Diagnosing and Predicting Individual Customer Defection in a Contractual Setting*, Singapore: Singapore Management University.

Borle, Sharad, Siddharth Singh, and Dipak Jain (2008), "Customer Lifetime Value Measurement," *Management Science*, 54 (1), 100–112.

Boucheron, Stéphane and Elisabeth Gassiat (2009), "A Bernstein-Von Mises Theorem for Discrete Probability Distributions," *Electronic Journal of Statistics*, 3, 114–148.

Boulding, William, Richard Staelin, Michael Ehret, and Wesley J. Johnston (2005), "A CRM Roadmap: What We Know, Potential Pitfalls and Where to Go," *Journal of Marketing*, 69 (4), 155–166.

Box, George E.P. and Gwilym M. Jenkins (1976), *Time Series Analysis: Forecasting and Control*, Oakland, CA: Holden-Day.

Braun, Michael and Paul Damien (2015), "Scalable Rejection Sampling for Bayesian Hierarchical Models," *Marketing Science*, forthcoming.

Breugelmans, Els, Tammo Bijmolt, Jie Zhang, Leonardo Basso, Matilda Dorotic, Praveen Kopalle, Alec Minnema, Willem Mijnlieff, and Nancy Wünderlich (2015), "Advancing Research on Loyalty Programs: A Future Research Agenda," *Marketing Letters*, 26 (2), 127 – 139.

Brody, Robert P. and Scott M. Cunningham (1968), "Personality Variables and the Consumer Decision Process," *Journal of Marketing Research*, 5 (1), 50–57.

Brooks, Stephen P. and Andrew Gelman (1998), "General Methods for Monitoring Convergence of Iterative Simulations," *Journal of Computational and Graphical Statistics*, 7 (4), 434–455.

Brooks, Steve, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng (2011), *Handbook of Markov Chain Monte Carlo*, Boca Raton, FL: Chapman & Hall/CRC.

Butler, Patrick and Joe Peppard (1998), "Consumer Purchasing on the Internet: Processes and Prospects," *European Management Journal*, 16 (5), 600–610.

Canova, Fabio (1993), "Forecasting Time Series with Common Seasonal Patterns," *Journal of Econometrics*, 55 (1), 173–200.

Carpenter, Bob, Daniel Lee, Marcus A. Brubaker, Allen Riddell, Andrew Gelman, Ben Goodrich, Jiqiang Guo, Matt Hoffman, Michael Betancourt, and Peter Li (2015), "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, forthcoming.

Chatfield, Christopher and Gerald J. Goodhardt (1973), "A Consumer Purchasing Model With Erlang Inter-Purchase Time," *Journal of the American Statistical Association*, 68 (344), 828–835.

Chib, Siddharta and Edward Greenberg (1996), "Markov Chain Monte Carlo Simulation Methods in Econometrics," *Econometric Theory*, 12 (3), 409–431.

Colombo, Richard and Weina Jiang (1999), "A Stochastic RFM Model," *Journal of Interactive Marketing*, 13 (3), 2–12.

Congdon, Peter (2003), *Applied Bayesian Modelling*, West Sussex: Wiley.

Corberán-Vallet, Ana, José D. Bermúdez, and Enriqueta Vercher (2011), "Forecasting Correlated Time Series With Exponential Smoothing Models," *International Journal of Forecasting*, 27 (2), 252–265.

Cowles, Mary Kathryn and Bradley P. Carlin (1996), "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review," *Journal of the American Statistical Association*, 91 (434), 883–904.

Cui, Geng, Man Leung Wong, and Hon-Kwong Lui (2006), "Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming," *Management Science*, 52 (4), 597–612.

Dagum, Estela Bee and Benoit Quenneville (1993), "Dynamic Linear Models for Time Series Components," *Journal of Econometrics*, 55 (1–2), 333–351.

Datta, Hannes, Bram Foubert, and Harald J. Van Heerde (2015), "The Challenge of Retaining Customers Acquired with Free Trials," *Journal of Marketing Research*, 52 (2), 217–234.

De Gooijer, Jan G. and Rob J. Hyndman (2006), "25 Years of Time Series Forecasting," *International Journal of Forecasting*, 22 (3), 443–473.

Dorotic, Matilda, Tammo H.A. Bijmolt, and Peter C. Verhoef (2012), "Loyalty Programmes: Current Knowledge and Research Directions," *International Journal of Management Reviews*, 14 (3), 217–237.

Dowling, Graham W. and Mark Uncles (1997), "Do Loyalty Programs Really Work?" *Sloan Management Review*, 38 (4), 71–82.

Dutta, Santanu (2015), "Local Smoothing for Kernel Distribution Function Estimation," *Communications in Statistics-Simulation and Computation*, 44 (4), 878–891.

Dwyer, Robert F. (1997), "Customer Lifetime Valuation to Support Marketing Decision Making," *Journal of Direct Marketing*, 11 (4), 6–13.

Ehrenberg, Andrew S.C. (1959), "The Pattern of Consumer Purchases," *Journal of the Royal Statistical Society. Series C*, 8 (1), 26–41.

——— (1988), *Repeat-Buying, Theory and Applications*, 2nd edition, London: Griffin.

Engle, Robert F. and Clive W.J. Granger (1987), "Co-Integration and Error Correction: Representation, Estimation, and Testing," *Econometrica*, 55 (2), 251–276.

Esteban-Bravo, Mercedes, Jose M. Vidal-Sanz, and Gökhan Yildirim (2014), "Valuing Customer Portfolios with Endogenous Mass and Direct Marketing Interventions Using a Stochastic Dynamic Programming Decomposition," *Marketing Science*, 33 (5), 621–640.

Evanschitzky, Heiner, Florian von Wangenheim, and Nancy V. Wünderlich (2012), "Perils of Managing the Service Profit Chain: The Role of Time Lags and Feedback Loops," *Journal of Retailing*, 88 (3), 356–366.

van Everdingen, Yvonne M., Wouter B. Aghina, and Dennis Fok (2005), "Forecasting Cross-Population Innovation Diffusion: A Bayesian Approach," *International Journal of Research in Marketing*, 22 (3), 293–308.

Fader, Peter S. and Bruce G.S. Hardie (2007), "How to Project Customer Retention," *Journal of Interactive Marketing*, 21 (1), 76–90.

——— (2009), "Probability Models for Customer-Base Analysis," *Journal of Interactive Marketing*, 23 (1), 61–69.

Fader, Peter S., Bruce G.S. Hardie, and Ka Lok Lee (2005a), "'Counting Your Customers' the Easy Way: An Alternative to the Pareto/NBD Model," *Marketing Science*, 24 (2), 275–285.

——— (2005b), "RFM and CLV: Using Iso-Value Curves for Customer Base Analysis," *Journal of Marketing Research*, 42 (4), 415–430.

Fader, Peter S., Bruce G.S. Hardie, and Jen Shang (2010), "Customer-Base Analysis in a Discrete-Time Noncontractual Setting," *Marketing Science*, 29 (6), 1086–1108.

Fawcett, Tom (2006), "An Introduction to ROC Analysis," *Pattern Recognition Letters*, 27 (8), 861–874.

Findley, David F., Brian C. Monsell, William R. Bell, Mark C. Otto, and Bor-Chung Chen (1998), "New Capabilities and Methods of the X-12-ARIMA Seasonal-Adjustment Program," *Journal of Business & Economic Statistics*, 16 (2), 127–152.

Fox, Justin (2015), "From 'Economic Man' to Behavioral Economics," *Harvard Business Review*, 93 (5), 78 – 85.

Franses, Philip Hans (1991), "Primary Demand for Beer in the Netherlands: An Application of ARMAX Model Specification," *Journal of Marketing Research*, 28 (2), 240–245.

——— (1994), "Modeling New Product Sales: An Application of Cointegration Analysis," *International Journal of Research in Marketing*, 11 (5), 491–502.

Gómez, Miguel I., Edward W. McLaughlin, and Dick R. Wittink (2004), "Customer Satisfaction and Retail Sales Performance: An Empirical Investigation," *Journal of Retailing*, 80 (4), 265–278.

Gönül, Füsun F. and Frenkel Ter Hofstede (2006), "How to Compute Optimal Catalog Mailing Decisions," *Marketing Science*, 25 (1), 65–74.

Gaukler, Gary M. (2010), "Preventing Avoidable Stockouts: The Impact of Item-Level RFID in Retail," *Journal of Business & Industrial Marketing*, 25 (8), 572–581.

Gedenk, Karen (2002), *Verkaufsförderung*, München: Vahlen.

Gefang, Deborah (2014), "Bayesian Doubly Adaptive Elastic-Net Lasso for VAR Shrinkage," *International Journal of Forecasting*, 30 (1), 1–11.

Gelfand, Alan E., Sujit K. Sahu, and Bradley P. Carlin (1995), "Efficient Parametrisations for Normal Linear Mixed Models," *Biometrika*, 82 (3), 479–488.

Gelfand, Alan E. and Adrian F.M. Smith (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85 (410), 398–409.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin (2003), *Bayesian Data Analysis*, 2nd edition, London: Chapman & Hall.

Gelman, Andrew, David A. van Dyk, Zaiying Huang, and W. John Boscardin (2009), “Using Redundant Parameterizations to Fit Hierarchical Models,” *Journal of Computational and Graphical Statistics*, 17 (1), 95–112.

Gelman, Andrew and Jennifer Hill (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, New York, NY: Cambridge University Press.

Gelman, Andrew and Donald B. Rubin (1992), “Inference from Iterative Simulation Using Multiple Sequences,” *Statistical Science*, 7 (4), 457–472.

Geman, Stuart and Donald Geman (1984), “Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6 (6), 721–741.

Germann, Frank, Gary L. Lilien, Lars Fiedler, and Matthias Kraus (2014), “Do Retailers Benefit from Deploying Customer Analytics?” *Journal of Retailing*, 90 (4), 587 – 593.

Geurts, Michael D. and I.B. Ibrahim (1975), “Comparing the Box-Jenkins Approach With the Exponentially Smoothed Forecasting Model Application to Hawaii Tourists,” *Journal of Marketing Research*, 12 (2), 182–188.

Gigerenzer, Gerd (2001), “The Adaptive Toolbox,” in *Bounded Rationality: The Adaptive Toolbox*, Gerd Gigerenzer and Reinhard Selten, eds., Cambridge, MA: MIT Press, 37–50.

——— (2006), “Follow the Leader,” *Harvard Business Review*, 84 (2), 58–59.

Gigerenzer, Gerd and Wolfgang Gaissmaier (2011), “Heuristic Decision Making,” *Annual Review of Psychology*, 62, 451–482.

Gigerenzer, Gerd and Daniel G. Goldstein (1996), “Reasoning the Fast and Frugal Way: Models of Bounded Rationality,” *Psychological Review*, 103 (4), 650–669.

Glady, Nicolas, Bart Baesens, and Christophe Croux (2009), “A Modified Pareto/NBD Approach for Predicting Customer Lifetime Value," *Expert Systems with Applications*, 36 (2), 2062–2071.

Glady, Nicolas, Aurélie Lemmens, and Christophe Croux (2015), "Unveiling the Relationship Between the Transaction Timing, Spending and Dropout Behavior of Customers," *International Journal of Research in Marketing*, 32 (1), 78 – 93.

Gneezy, Ayelet, Uri Gneezy, and Dominique Olié Lauga (2014), "A Reference-Dependent Model of the Price–Quality Heuristic," *Journal of Marketing Research*, 51 (2), 153–164.

Goldstein, Daniel G. and Gerd Gigerenzer (2009), "Fast and Frugal Forecasting," *International Journal of Forecasting*, 25 (4), 760–772.

Goodman, David A. and Kavin W. Moody (1970), "Determining Optimum Price Promotion Quantities," *Journal of Marketing*, 34 (4), 31–39.

Green, Peter J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Greene, William H. (2008), *Econometric Analysis*, 6th edition, Upper Saddle River, NJ: Prentice Hall.

Guo, Ruey-Shan (2009), "A Multi-Category Inter-Purchase Time Model Based on Hierarchical Bayesian Theory," *Expert Systems with Applications*, 36 (3), 6301–6308.

Gupta, Sunil and Valarie A. Zeithaml (2006), "Customer Metrics and Their Impact on Financial Performance," *Marketing Science*, 25 (6), 718–739.

Haenlein, Michael, Andreas M. Kaplan, and Anemone J. Beeser (2007), "A Model to Determine Customer Lifetime Value in a Retail Banking Context," *European Management Journal*, 25 (3), 221–234.

Haenlein, Michael, Andreas M. Kaplan, and Detlef Schoder (2006), "Valuing the Real Option of Abandoning Unprofitable Customers When Calculating Customer Lifetime Value," *Journal of Marketing*, 70 (3), 5–20.

Hagerty, Michael R. (1987), "Conditions Under Which Econometric Models Will

Outperform Naive Models," *International Journal of Forecasting*, 3 (3–4), 457 – 460.

Hardie, Bruce G.S., Eric J. Johnson, and Peter S. Fader (1993), "Modeling Loss Aversion and Reference Dependence Effects on Brand Choice," *Marketing Science*, 12 (4), 378–394.

Hastings, W. Keith (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57 (1), 97–109.

Heskett, James L., Thomas O. Jones, Gary W. Loveman, W. Earl Sasser Jr., and Leonard A. Schlesinger (1994), "Putting the Service-Profit Chain to Work," *Harvard Business Review*, 72 (2), 164–174.

Ho, Teck-Hua, Young-Hoon Park, and Yong-Pin Zhou (2006), "Incorporating Satisfaction into Customer Value Analysis: Optimal Investment in Life-Time Value," *Marketing Science*, 25 (3), 260–277.

Holak, Susan L. and Y.E. Tang (1990), "Advertising's Effect on the Product Evolutionary Cycle," *Journal of Marketing*, 54 (3), 16–29.

Holt, Charles C. (2004a), "Author's Retrospective on 'Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages'," *International Journal of Forecasting*, 20 (1), 11–13.

——— (2004b), "Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages," *Journal of Economic & Social Measurement*, 29 (1-3), 123–125.

——— (2004c), "Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages," *International Journal of Forecasting*, 20 (1), 5–10.

Homan, Matthew D. and Andrew Gelman (2014), "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo," *The Journal of Machine Learning Research*, 15 (1), 1593–1623.

Hu, Ya-Han and Tzu-Wei Yeh (2014), "Discovering Valuable Frequent Patterns Based on RFM Analysis Without Customer Identification Information," *Knowledge-Based Systems*, 61, 76–88.

Huang, Chun-Yao (2012), "To Model, or Not to Model: Forecasting for Customer Prioritization," *International Journal of Forecasting*, 28 (2), 497–506.

Hughes, Arthur (1996), *The Complete Database Marketer: Second Generation Strategies and Techniques for Tapping the Power of Your Customer Database*, 2nd edition, New York, NY: McGraw-Hill.

Hyndman, Rob J. (2014), "Forecast Package for R," Available from: `http://github.com/robjhyndman/forecast` [accessed 6/2/2014].

Hyndman, Rob J., Roman A. Ahmed, George Athanasopoulos, and Han Lin Shang (2011), "Optimal Combination Forecasts for Hierarchical Time Series," *Computational Statistics & Data Analysis*, 55 (9), 2579–2589.

Hyndman, Rob J. and Yeasmin Khandakar (2008), "Automatic Time Series Forecasting: The Forecast Package for R," *Journal of Statistical Software*, 26 (3), 1–22.

Hyndman, Rob J. and Anne B. Koehler (2006), "Another Look at Measures of Forecast Accuracy," *International Journal of Forecasting*, 22 (4), 679–688.

Hyndman, Rob J., Anne B. Koehler, Ralph D. Snyder, and Simone Grose (2002), "A State Space Framework for Automatic Forecasting Using Exponential Smoothing Methods," *International Journal of Forecasting*, 18 (3), 439–454.

Jackson, Barbara (1985), *Winning and Keeping Industrial Customers*, Lexington, MA: D.C. Heath and Company.

Jackson, Robert R. and Paul Wang (1996), *Strategic Database Marketing*, Lincolnwod, IL: NTC.

Jain, Dipak and Siddharth Singh (2002), "Customer Lifetime Value Research in Marketing: A Review and Future Directions," *Journal of Interactive Marketing*, 16 (2), 34–47.

Jen, Lichung, Chien-Heng Chou, and Greg M. Allenby (2009), "The Importance of Modeling Temporal Dependence of Timing and Quantity in Direct Marketing," *Journal of Marketing Research*, 46 (4), 482–493.

Jerath, Kinshuk, Peter S. Fader, and Bruce G.S. Hardie (2011), "New Perspectives on Customer "Death" Using a Generalization of the Pareto/NBD Model," *Marketing Science*, 30 (5), 866–880.

Kao, Yu-Ting, Hsin-Hung Wu, Hsuan-Kai Chen, and En-Chi Chang (2011), "A Case Study of Applying LRFM Model and Clustering Techniques to Evaluate Customer Values," *Journal of Statistics and Management Systems*, 14 (2), 267–276.

Keiningham, Timothy L., Bruce Cooil, Lerzan Aksoy, Tor W. Andreassen, and Jay Weiner (2007), "The Value of Different Customer Satisfaction and Loyalty Metrics in Predicting Customer Retention, Recommendation, and Share-of-Wallet," *Managing Service Quality*, 17 (4), 361 – 384.

Khan, Romana, Michael Lewis, and Vishal Singh (2009), "Dynamic Customer Management and the Value of One-to-One Marketing," *Marketing Science*, 28 (6), 1063–1079.

Kleiber, Christian and Samuel Kotz (2003), *Statistical Size Distributions in Economics and Actuarial Sciences*, West Sussex: Wiley.

Krasnikov, Alexander, Satish Jayachandran, and V. Kumar (2009), "The Impact of Customer Relationship Management Implementation on Cost and Profit Efficiencies: Evidence from the U.S. Commercial Banking Industry," *Journal of Marketing*, 73 (6), 61 – 76.

Kumar, V. (2013), *Profitable Customer Engagement: Concept, Metrics, and Strategies*, Thousand Oaks, CA: Sage.

Kumar, V., Yashoda Bhagwat, and Xi Zhang (2015), "Regaining 'Lost' Customers: The Predictive Power of First-Lifetime Behavior, the Reason for Defection, and the Nature of the Win-Back Offer," *Journal of Marketing*, 79 (4), 34–55.

Kumar, V., J. Andrew Petersen, and Robert P. Leone (2010), "Driving Profitability by Encouraging Customer Referrals: Who, When, and How," *Journal of Marketing*, 74 (3), 1–17.

Kumar, V. and Werner J. Reinartz (2012), *Customer Relationship Management: Concept, Strategy, and Tools*, Heidelberg: Springer.

Kumar, V., Denish Shah, and Rajkumar Venkatesan (2006), "Managing Retailer Profitability - One Customer at a Time!" *Journal of Retailing*, 82 (4), 277–294.

Kumar, Vineet, Kannan Srinivasan, Vithala R. Rao, Yao Zhang, Eric T. Bradlow, and Dylan S. Small (2015), "Commentaries and Reply on "Predicting Customer Value Using Clumpiness: From RFM to RFMC" by Yao Zhang, Eric T. Bradlow, and Dylan S. Small," *Marketing Science*, 34 (2), 209–217.

Lamb, Charles W., Joe F. Hair, and Carl D. McDaniel (2008), *Marketing*, 10th edition, Mason, OH: South-Western/Cengage.

Larivière, Bart (2008), "Linking Perceptual and Behavioral Customer Metrics to Multiperiod Customer Profitability," *Journal of Service Research*, 11 (1), 3–21.

Larivière, Bart, Timothy L. Keiningham, Lerzan Aksoy, Atakan Yalçin, Forrest V. Morgeson III, and Sunil Mithas (2015), "Modeling Heterogeneity in The Satisfaction, Loyalty Intention and Shareholder Value Linkage: A Cross-Industry Analysis at the Customer and Firm Level," *Journal of Marketing Research*, forthcoming.

Larivière, Bart and Dirk Van den Poel (2007), "Banking Behaviour After the Lifecycle Event of 'Moving in Together': An Exploratory Study of the Role of Marketing Investments," *European Journal of Operational Research*, 183 (1), 345–369.

Law, Agnes K.Y., Christine T. Ennew, and Darryn Mitussis (2013), "Adoption of Customer Relationship Management in the Service Sector and Its Impact on Performance," *Journal of Relationship Marketing*, 12 (4), 301–330.

Lee, Michael D. and Tarrant D.R. Cummins (2004), "Evidence Accumulation in Decision Making: Unifying the 'Take the Best' and the 'Rational' Models," *Psychonomic Bulletin & Review*, 11 (2), 343–352.

Leeflang, Peter S.H., Philippe A. Naert, Michael Wedel, and Dick R. Wittink (2000), *Building Models for Marketing Decisions*, Berlin: Springer.

Lemon, Katherine N. and Florian von Wangenheim (2009), "The Reinforcing Effects of Loyalty Program Partnerships and Core Service Usage: A Longitudinal Analysis," *Journal of Service Research*, 11 (4), 357–370.

Lenk, Peter and Bryan Orme (2009), "The Value of Informative Priors in Bayesian Inference with Sparse Data," *Journal of Marketing Research*, 46 (6), 832–845.

Lichung, Jen, Chou Chien-Heng, and Greg M. Allenby (2003), "A Bayesian Approach to Modeling Purchase Frequency," *Marketing Letters*, 14 (1), 5–20.

Liu, Yuping and Rong Yang (2009), "Competing Loyalty Programs: Impact of Market Saturation, Market Share, and Category Expandability," *Journal of Marketing*, 73 (1), 93–108.

Lovell, Michael C. (1963), "Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis," *Journal of the American Statistical Association*, 58 (304), 993–1010.

Lunn, David, Christopher Jackson, Nicky Best, Andrew Thomas, and David Spiegelhalter (2012), *The BUGS Book: A Practical Introduction to Bayesian Analysis*, Boca Raton, FL: Chapman & Hall/CRC.

Lunn, David J., Andrew Thomas, Nicky Best, and David Spiegelhalter (2000), "WinBUGS - A Bayesian Modelling Framework: Concepts, Structure, and Extensibility," *Statistics and Computing*, 10 (4), 325–337.

Lysonski, Steven and Srinivas Durvasula (2013), "Consumer Decision Making Styles in Retailing: Evolution of Mindsets and Psychological Impacts," *Journal of Consumer Marketing*, 30 (1), 75–87.

Ma, Shaohui, Hui Tan, and Fang Shu (2015), "When Is the Best Time to Reactivate Your Inactive Customers?" *Marketing Letters*, 26 (1), 81–98.

Makridakis, Spyros and Michèle Hibon (2000), "The M3-Competition: Results, Conclusions and Implications," *International Journal of Forecasting*, 16 (4), 451–476.

Makridakis, Spyros and Steven C. Wheelwright (1977), "Forecasting: Issues and Challenges for Marketing Management," *Journal of Marketing*, 41 (4), 24–38.

Manchanda, Puneet, Jean-Pierre Dubé, Khim Yong Goh, and Pradeep K. Chintagunta (2006), "The Effect of Banner Advertising on Internet Purchasing," *Journal of Marketing Research*, 43 (1), 98–108.

Marinova, Detelina and Jagdip Singh (2014), "Consumer Decision to Upgrade or Downgrade a Service Membership," *Journal of the Academy of Marketing Science*, 42 (6), 596–618.

Mark, Tanya, Katherine N. Lemon, Mark Vandenbosch, Jan Bulla, and Antonello Maruotti (2013), "Capturing the Evolution of Customer–Firm Relationships: How Customers Become More (or Less) Valuable Over Time," *Journal of Retailing*, 89 (3), 231–245.

Marshall, Pablo (2015), "A Simple Heuristic for Obtaining Pareto/NBD Parameter Estimates," *Marketing Letters*, 26 (2), 165–173.

McCall, Michael and Clay Voorhees (2010), "The Drivers of Loyalty Program Success: An Organizing Framework and Research Agenda," *Cornell Hospitality Quarterly*, 51 (1), 35–52.

McLaughlin, Robert L. (1963), "The Breakthrough in Sales Forecasting," *Journal of Marketing*, 27 (2), 46–54.

Meade, Nigel and Towhidul Islam (2010), "Using Copulas to Model Repeat Purchase Behaviour - An Exploratory Analysis via a Case Study," *European Journal of Operational Research*, 200 (3), 908–917.

Mentzer, John T. (1988), "Forecasting With Adaptive Extended Exponential Smoothing," *Journal of the Academy of Marketing Science*, 16 (3-4), 62–70.

Mentzer, John T. and Roger Gomes (1994), "Further Extensions of Adaptive Extended Exponential Smoothing and Comparison With the M-Competition," *Journal of the Academy of Marketing Science*, 22 (4), 372–382.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller (1953), "Equation of State Calculations by Fast Computing Machines," *The Journal of Chemical Physics*, 21 (6), 1087–1092.

Mithas, Sunil, M.S. Krishnan, and Claes Fornell (2005), "Why Do Customer Relationship Management Applications Affect Customer Satisfaction?" *Journal of Marketing*, 69 (4), 201–209.

Montaguti, Elisa, Scott A. Neslin, and Sara Valentini (2015), "Can Marketing Campaigns Induce Multichannel Buying and More Profitable Customers? A Field Experiment," *Marketing Science*, forthcoming.

Moriarty, Mark and Gerald Salamon (1980), "Estimation and Forecast Performance of a Multivariate Time Series Model of Sales," *Journal of Marketing Research*, 17 (4), 558–564.

Mowen, John C. (1988), "Beyond Consumer Decision Making," *Journal of Consumer Marketing*, 5 (1), 15–25.

Mulhern, Frank (1999), "Customer Profitability Analysis: Measurement, Concentration, and Research Directions," *Journal of Interactive Marketing*, 13 (1), 25–40.

Mzoughia, Mohamed Ben and Mohamed Limam (2014), "An Improved BG/NBD Approach for Modeling Purchasing Behavior Using COM-Poisson Distribution," *International Journal of Modeling and Optimization*, 4 (2), 141.

Neal, Radford M. (1993), "Probabilistic Inference Using Markov Chain Monte Carlo Methods," Available from: `http://omega.albany.edu:8008/neal.pdf` [accessed 9/20/2013].

——— (1997), "Markov Chain Monte Carlo Methods Based on 'Slicing' the Density Function, Technical Report No. 9722, Dept. of Statistics, University of Toronto," Available from: `http://www.cs.utoronto.ca/~radford/ftp/slice.pdf` [accessed 10/13/2013].

Neslin, Scott, Gail Taylor, Kimberly Grantham, and Kimberly McNeil (2013), "Overcoming the 'Recency Trap' in Customer Relationship Management," *Journal of the Academy of Marketing Science*, 41 (3), 320 – 337.

Newell, Ben R., Nicola J. Weston, and David R. Shanks (2003), "Empirical Tests of a Fast-and-Frugal Heuristic: Not Everyone 'Takes-the-Best'," *Organizational Behavior & Human Decision Processes*, 91 (1), 82–96.

Nikumanesh, Esmaeil and Amir Albadvi (2014), "Customer's Life–Time Value Using the RFM Model in the Banking Industry: A Case Study," *International Journal of Electronic Customer Relationship Management*, 8 (1-3), 15–30.

Niu, Han-Jen (2013), "Cyber Peers' Influence for Adolescent Consumer in Decision-Making Styles and Online Purchasing Behavior," *Journal of Applied Social Psychology*, 43 (6), 1228–1237.

Ntzoufras, Ioannis (2009), *Bayesian Modeling Using WinBUGS*, West Sussex: Wiley.

van Oest, Rutger and George Knox (2011), "Extending the BG/NBD: A Simple Model of Purchases and Complaints," *International Journal of Research in Marketing*, 28 (1), 30–37.

Oliver, Richard L. (1997), *Satisfaction: A Behavioral Perspective on the Consumer*, Boston, MA: Irwin McGraw-Hill.

——— (1999), "Whence Customer Loyalty?" *Journal of Marketing*, 63 (4), 33–44.

Ord, Keith (2004), "Charles Holt's Report on Exponentially Weighted Moving Averages: An Introduction and Appreciation," *International Journal of Forecasting*, 20 (1), 1–3.

Park, Sungho and Sachin Gupta (2011), "A Regime-Switching Model of Cyclical Category Buying," *Marketing Science*, 30 (3), 469–480.

Payne, Adrian and Pennie Frow (2005), "A Strategic Framework for Customer Relationship Management," *Journal of Marketing*, 69 (4), 167–176.

Peppers, Don and Martha Rogers (1993), *The One-to-One Future: Building Relationships One Customer at a Time*, New York, NY: Doubleday.

Petersen, J. Andrew, Leigh McAlister, David J. Reibstein, Russell S. Winer, V. Kumar, and Geoff Atkinson (2009), "Choosing the Right Metrics to Maximize Profitability and Shareholder Value," *Journal of Retailing*, 85 (1), 95–111.

Plummer, Martyn (2003), "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna/Austria*, 1–10.

Radas, Sonja and Steven M. Shugan (1998), "Seasonal Marketing and Timing New Product Introductions," *Journal of Marketing Research*, 35 (3), 296–315.

Raftery, Adrian E. and Steven Lewis (1992), "How Many Iterations in the Gibbs Sampler?" in *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Conference on Bayesian Statistics*, José M. Bernando, James O. Berger, A. Philip Dawid, and Adrian F.M. Smith, eds., Oxford: Oxford University Press, 763–773.

Raynauld, Jacques and Jean-Guy Simonato (1993), "Seasonal BVAR Models: A Search Along Some Time Domain Priors," *Journal of Econometrics*, 55 (1–2), 203–229.

Reichheld, Frederick F., Robert G. Markey Jr., and Christopher Hopton (2000), "The Loyalty Effect - The Relationship Between Loyalty and Profits," *European Business Journal*, 12 (3), 134–139.

Reichheld, Frederick F. and W. Earl Jr. Sasser (1991), "Zero-Migration: Dienstleister im Sog der Qualitätsrevolution," *Harvard Business Manager*, 13 (4), 108–116.

Reichheld, Frederick F. and W. Earl Sasser Jr. (1990a), "Zero Defections: Quality Comes to Services," *Harvard Business Review*, 68 (5), 105–111.

——— (1990b), "Zero defections: Quality comes to services." *Harvard Business Review*, 68 (5), 105 – 111.

Reichheld, Frederick F. and Thomas Teal (1996), *The Loyalty Effect: The Hidden Force Behind Growth, Profits, and Lasting Value*, Boston, MA: Harvard Business School Press.

Reinartz, Werner J. and V. Kumar (2000), "On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing," *Journal of Marketing*, 64 (4), 17–35.

——— (2002), "The Mismanagement of Customer Loyality," *Harvard Business Review*, 80 (7), 86–94.

——— (2003), "The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration," *Journal of Marketing*, 67 (January), 77–99.

Reinartz, Werner J., Jacquelyn S. Thomas, and V. Kumar (2005), "Balancing Acquisition and Retention Resources to Maximize Customer Profitability," *Journal of Marketing*, 69 (1), 63–79.

Reutterer, Thomas, Andreas Mild, Martin Natter, and Alfred Taudes (2006), "A Dynamic Segmentation Approach for Targeting and Customizing Direct Marketing Campaigns," *Journal of Interactive Marketing*, 20 (3/4), 43–57.

Rieskamp, Jörg and Ulrich Hoffrage (2008), "Inference Under Time Pressure: How Opportunity Costs Affect Strategy Selection," *Acta Psychologica*, 127 (2), 258–276.

Roberts, Gareth O. and Sujit K. Sahu (1997), "Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler," *Journal of the Royal Statistical Society. Series B*, 59 (2), 291–317.

Roberts, Mary Lou and Paul D. Berger (1999), *Direct Marketing Management*, 2nd edition, Upper Saddle River, NJ: Prentice Hall.

Roel, Raymond (1988), "Direct Marketing's 50 Big Ideas," *Direct Marketing*, 50 (May), 45–52.

Romero, Jaime, Ralf Van der Lans, and Berend Wierenga (2013), "A Partially Hidden Markov Model of Customer Dynamics for CLV Measurement," *Journal of Interactive Marketing*, 27 (3), 185–208.

Rossi, Peter E., Greg M. Allenby, and Rob McCulloch (2005), *Bayesian Statistics and Marketing*, New York, NY: John Wiley and Sons.

Rust, Roland, Katherine N. Lemon, and Valarie A. Zeithaml (2004), "Return on

Marketing: Using Customer Equity to Focus Marketing Strategy," *Journal of Marketing*, 68 (1), 109–127.

Rust, Roland T., Valarie A. Zeithaml, and Katherine N. Lemon (2000), *Driving Customer Equity: How Customer Lifetime Value is Reshaping Corporate Strategy*, New York, NY: The Free Press.

Ryals, Lynette (2005), "Making Customer Relationship Management Work:The Measurement and Profitable Management of Customer Relationships," *Journal of Marketing*, 69 (4), 252–261.

Ryals, Lynette and Adrian Payne (2001), "Customer Relationship Management in Financial Services: Towards Information-Enabled Relationship Marketing," *Journal of Strategic Marketing*, 9 (1), 3–27.

Sackmann, Stefan, Dennis Kundisch, and Markus Ruch (2010), "Customer Portfolio Management in E-Commerce: An Analytical Model for Optimization," *Management Research Review*, 33 (6), 617–634.

Scherer, Anne, Nancy V. Wünderlich, and Florian von Wangenheim (2015), "The Value of Self-Service: Long-Term Effects of Technology-Based Self-Service Usage on Customer Retention," *Management Information Systems Quarterly*, 39 (1), 177–200.

Schliep, Erin M. and Jennifer A. Hoeting (2015), "Data Augmentation and Parameter Expansion for Independent or Spatially Correlated Ordinal Data," *Computational Statistics & Data Analysis*, 90, 1 – 14.

Schmittlein, David C., Donald G. Morrison, and Richard Colombo (1987), "Counting Your Customers: Who Are They And What Will They Do Next?" *Management Science*, 33 (1), 1–24.

Schmittlein, David C. and Robert A. Peterson (1994), "Customer Base Analysis: An Industrial Purchase Process Application," *Marketing Science*, 13 (1), 41–67.

Schweidel, David A., Peter S. Fader, and Eric T. Bradlow (2008), "Understanding Service Retention Within and Across Cohorts Using Limited Information," *Journal of Marketing*, 72 (1), 82–94.

Schweidel, David A. and George Knox (2013), "Incorporating Direct Marketing Activity into Latent Attrition Models," *Marketing Science*, 32 (3), 471–487.

Schweidel, David A., Young-Hoon Park, and Zainab Jamal (2014), "A Multiactivity Latent Attrition Model for Customer Base Analysis," *Marketing Science*, 33 (2), 273–286.

Semon, Thomas T. (1958), "A Simple Seasonal Pattern," *Journal of Marketing*, 23 (1), 62–63.

Sevilla, Julio and Barbara E. Kahn (2014), "The Completeness Heuristic: Product Shape Completeness Influences Size Perceptions, Preference, and Consumption," *Journal of Marketing Research*, 51 (1), 57–68.

Shah, Anuj K. and Daniel M. Oppenheimer (2008), "Heuristics Made Easy: An Effort-Reduction Framework," *Psychological Bulletin*, 134 (2), 207.

Shah, Denish, V. Kumar, and Kihyun Hannah Kim (2014), "Managing Customer Profits: The Power of Habits," *Journal of Marketing Research*, 51 (6), 726–741.

Shu-Chuan, Lo (2008), "Online Customer Identification Based on Bayesian Model of Interpurchase Times and Recency," *International Journal of Systems Science*, 39 (8), 853–863.

Shugan, Steven M. (2005), "Brand Loyalty Programs: Are They Shams?" *Marketing Science*, 24 (2), 185–193.

Solomon, Michael R. (2011), *Consumer Behavior*, 9th edition, Upper Saddle River, NJ: Pearson.

Soysal, Gonca P. and Lakshman Krishnamurthi (2012), "Demand Dynamics in the Seasonal Goods Industry: An Empirical Analysis," *Marketing Science*, 31 (2), 293–316.

Sproles, George B. and Elizabeth L. Kendall (1986), "A Methodology for Profiling Consumers' Decision-Marking Styles," *Journal of Consumer Affairs*, 20 (2), 267–279.

Stan, Development Team (2015), "Stan: A C++ Library for Probability and Sampling, Version 2.6.0," Available from: `http://mc-stan.org/` [accessed 2/19/2015].

Steffes, Erin M., B.P.S. Murthi, and Ram C. Rao (2008), "Acquisition, Affinity and Rewards: Do They Stay or Do They Go?" *Journal of Financial Services Marketing*, 13 (3), 221–233.

Strang, Kenneth David (2012), "Prioritization and Supply Chain Logistics as a Marketing Function in a Mining Company," *Journal of Marketing Channels*, 19 (2), 141–155.

Takada, Hirokazu and Frank M. Bass (1998), "Multiple Time Series Analysis of Competitive Marketing Behavior," *Journal of Business Research*, 43 (2), 97–107.

Taylor, James W. (2003), "Exponential Smoothing With a Damped Multiplicative Trend," *International Journal of Forecasting*, 19 (4), 715–725.

Thaivanich, Pattana, Rajesh K. Chandy, and Gerard J. Tellis (2000), "Which Ad Works, When, Where, and How Often? Modeling the Effects of Direct Television Advertising," *Journal of Marketing Research*, 37 (1), 32–46.

Trinh, Giang, Cam Rungie, Malcolm Wright, Carl Driesener, and John Dawes (2014), "Predicting Future Purchases with the Poisson Log-Normal Model," *Marketing Letters*, 25 (2), 219 – 234.

Venkatesan, Rajkumar and V. Kumar (2004), "A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy," *Journal of Marketing*, 68 (4), 106–125.

Venkatesan, Rajkumar, V. Kumar, and Timothy Bohling (2007), "Optimal Customer Relationship Management Using Bayesian Decision Theory: An Application for Customer Selection," *Journal of Marketing Research*, 44 (4), 579–594.

Wade, Betsy (1988), "PRACTICAL TRAVELER; Mileage Points Can Fly Away," Available from: `http://www.nytimes.com/1988/11/20/travel/`

`practical-traveler-mileage-points-can-fly-away.html` [accessed 14/07/2015].

Wünderlich, Robin (2010), *Prediction of Purchase Behavior: Benchmarking a Hierarchical Bayesian Model against Managerial Heuristics*, Master's thesis, Technische Universität München.

von Wangenheim, Florian and Tomás Bayón (2007), "The Chain from Customer Satisfaction via Word-of-Mouth Referrals to New Customer Acquisition," *Journal of the Academy of Marketing Science*, 35 (2), 233–249.

Wildt, Albert R. (1977), "Estimating Models of Seasonal Market Response Using Dummy Variables," *Journal of Marketing Research*, 14 (1), 34–41.

Winer, Russell S. (2001), "A Framework for Customer Relationship Management," *California Management Review*, 42 (4), 89–105.

Winters, Peter R. (1960), "Forecasting Sales by Exponentially Weighted Moving Averages," *Management Science*, 6 (3), 324–342.

Withycombe, Richard (1989), "Forecasting With Combined Seasonal Indices," *International Journal of Forecasting*, 5 (4), 547–552.

Wu, Couchen and Hsiu-Li Chen (2000a), "A Consumer Purchasing Model with Learning and Departure Behaviour," *The Journal of the Operational Research Society*, 51 (5), 583–591.

——— (2000b), "Counting Your Customers: Compounding Customer's In-Store Decisions, Interpurchase Time and Repurchasing Behavior," *European Journal of Operational Research*, 127 (1), 109–119.

Wübben, Markus and Florian von Wangenheim (2008), "Instant Customer Base Analysis: Managerial Heuristics Often 'Get It Right'," *Journal of Marketing*, 72 (3), 82–93.

Xiao, Sai, Athanasios Kottas, and Bruno Sansó (2015), "Modeling for Seasonal Marked Point Processes: An Analysis of Evolving Hurricane Occurrences," *The Annals of Applied Statistics*, 9 (1), 353–382.

Xie, Jinhong and Steven M. Shugan (2001), "Electronic Tickets, Smart Cards, and Online Prepayments: When and How to Advance Sell," *Marketing Science*, 20 (3), 219–243.

Xie, Jinhong, X. Michael Song, Marvin Sirbu, and Qiong Wang (1997), "Kalman Filter Estimation of New Product Diffusion Models," *Journal of Marketing Research*, 34 (3), 378–393.

Yi, Youjae (1990), "A Critical Review of Consumer Satisfaction," in *Review of Marketing*, Valarie A. Zeithaml, ed., Chicago, IL: AMA, 68–123.

Zeithaml, Valarie A. (1999), "Service Quality, Profitability and the Economic Worth of Customers: What We Know and What We Need to Learn," *Journal of the Academy of Marketing Science*, 28 (1), 67–85.

Zeithaml, Valarie A., Leonard L. Berry, and A. Parasuraman (1996), "The Behavioral Consequences of Service Quality," *Journal of Marketing*, 60 (2), 31–46.

Zeithaml, Valarie A., Ruth N. Bolton, John Deighton, Timothy L. Keiningham, Katherine N. Lemon, and J. Andrew Petersen (2006), "Forward-Looking Focus: Can Firms Have Adaptive Foresight?" *Journal of Service Research*, 9 (2), 168–183.

Zhang, Yao, Eric T. Bradlow, and Dylan S. Small (2015), "Predicting Customer Value Using Clumpiness: From RFM to RFMC," *Marketing Science*, 34 (2), 195–208.

Zinkhan, George M. (1992), "From the Editor: Human Nature and Models of Consumer Decision Making," *Journal of Advertising*, 4, ii – iii.

Zitzlsperger, David F.S., Thomas Robbert, and Stefan Roth (2009), "Forecasting Customer Buying Behaviour 'Controlling for Seasonality'," in *Proceedings of the ANZMAC Conference 2009, Dunedin/New Zealand*.

Appendix A

DMPT: Implementation, Program Code and Additional Figures

A.1 Implementation Details and BUGS Model Code

The source code to specify the MCMC simulator in the *WinBUGS* language for the DMPT is shown in Source Code Clipping A.1. The source code covers the full model and includes the hierarchical random-effects generalized gamma individual purchase model, the cross-sectional inverse generalized gamma model, the generalized gamma component mixture, the temporal dynamics submodel, and all prior definitions. The acronym *BUGS* stands for Bayesian Inference Using Gibbs sampling. The *BUGS* family of software packages allows for fitting arbitrarily complex Bayesian models and producing reliable Bayesian statistics for a wide range of statistical models using MCMC methods (Lunn et al. 2000).

The latinized names of the symbols are used in the source code clippings, e.g., α becomes *alpha*. Subscripts or additional identifiers are separated by a point, e.g., the precision for θ in denoted *prec.theta*. The variable N denotes the number of customers in the dataset and K is set to the number of mixture components. The matrix *ipt[i,j]* contains the interpurchase times for each customer *i*. The number of observed interpurchase times for customer *i* is stored in *len[i]*.

The heterogeneity across customers i is modeled so that the individual λ_i's follow an inverse generalized gamma distribution IGG(ν, θ, γ), as shown in Equation (6.3, p. 75). While *WinBUGS* provides constructs for the generalized gamma distribution (gen.gamma) and the standard gamma distribution (dgamma), it does not feature a construct for the inverse generalized gamma distribution. The definitions in *WinBUGS* are as follows (Lunn et al. 2000):

$$\text{gen.gamma}(r, \mu, \beta) = \frac{\beta}{\Gamma(r)} \mu^{\beta r} x^{\beta r - 1} e^{-(x\mu)^{\beta}} \tag{A.1}$$

$$\text{dgamma}(r, \mu) = \frac{\mu x^{r-1} e^{-\mu x}}{\Gamma(r)} \tag{A.2}$$

The gen.gamma construct does not allow for a negative β parameter. Therefore, a reparameterization in the form of IGG(r, μ, β) = gen.gamma ($r, \mu, -\beta$) is not feasible in *WinBUGS*. Fortunately, one can express the IGG in terms of the standard gamma distribution as follows (Kleiber and Kotz 2003, p. 148):

$$\lambda_i \sim \text{IGG}(\nu, \theta, \gamma) \tag{A.3}$$

$$\lambda_i^{-\gamma} \sim \text{dgamma}(\nu, \theta^{-\gamma}) \tag{A.4}$$

The parameterization of the generalized gamma distribution used in *WinBUGS* differs from the one used by Allenby, Leone, and Jen (1999), which is repeated here for readability:

$$t_{ij} \sim GG(\alpha, \lambda_i, \gamma) \overset{\text{p.d.f.}}{\rightleftharpoons} \frac{\gamma}{\Gamma(\alpha)\lambda_i^{\alpha\gamma}} t_{ij}^{\alpha\gamma - 1} e^{-(t_{ij}/\lambda_i)^{\gamma}} \tag{A.5}$$

While the parameters r and α as well as the parameters β and γ are equivalent, *WinBUGS* uses "precision" parametrization on the scale parameter μ instead of a variance parameter λ_i. This implies that $\mu = \lambda_i^{-1}$. I use a "logical node" that transforms the parameter accordingly. Note that *WinBUGS* parameterizes the multivariate normal distribution with a precision matrix instead of the usual variance-covariance matrix.

The prior on the variance-covariance matrix V is distributed inverted Wishart with the matrix G and the g as parameters. *WinBUGS* provides the Wishart distribution, so I use that that if $V \sim IW(G,g)$ then $V^{-1} \sim W(G^{-1},g)$. In addition *WinBUGS* expects the parameter matrix for the multivariate normal distribution and the Wishart distribution to be in precision format. With this in mind I translated the original prior $V \sim IW(G,g)$ to the *WinBUGS* statement: prec.V[,]~dwish$(G[,],g)$.

The parameters $a_0 = 10$, $b_0 = 10$, $g = 15$, $G = 15\mathbf{I}$, $\gamma_1 = 1.0$, $\gamma_2 = 1.4$, $\gamma_3 = 0.8$ and beta.prior= $\vec{0}$ are chosen as in Allenby, Leone, and Jen (1999).

Listing A.1: The DMPT Program Code in WinBUGS for $K = 3$ Components

```
model{
for (c in 1:K) {
    iggp.theta[c] <- pow (theta[c],-gamma[c])
    alpha[c] ~ dunif(0,50)
    nu[c] ~ dunif(0,50)
    }
for (i in 1:N) {
    for (c in 1:K) {
        igg.lambda[i,c] ~ dgamma (nu[c], iggp.theta[c])
        lambda[i,c] <- pow (igg.lambda[i,c], -1/gamma[c])
        prec.lambda[i,c] <- 1/lambda[i,c]
        }
    beta[i,1:5]~ dmnorm (bar.beta[], prec.V[,])
    for (j in 4:len[i]) {
        k.phix[i,j]<-phi(log(ipt[i,j-1])*beta[i,3]
                        +log(ipt[i,j-2])*beta[i,4]
                        +log(ipt[i,j-3])*beta[i,5]
                        +beta[i,2])
        p[i,j,1] <- 1-phi(beta[i,1])
        p[i,j,2] <- phi(beta[i,1])*(1-k.phix[i,j])
        p[i,j,3] <- phi(beta[i,1])*(k.phix[i,j])
        k[i,j]   ~  dcat (p[i,j,])
        ipt[i,j] ~  gen.gamma (alpha[k[i,j]],
                               prec.lambda[i,k[i,j]],
                               gamma[k[i,j]])
        }
    }
igg.b01<-pow(b0,-gamma[1])
igg.theta[1]~dgamma(a0,igg.b01)I(,igg.theta[2])
theta[1]<-pow(igg.theta[1],-1/gamma[1])

igg.b02<-pow(b0,-gamma[2])
igg.theta[2]~dgamma(a0,igg.b02)I(igg.theta[1],igg.theta[3])
theta[2]<-pow(igg.theta[2],-1/gamma[2])

igg.b03<-pow(b0,-gamma[3])
igg.theta[3]~dgamma(a0,igg.b03)I(igg.theta[2],)
theta[3]<-pow(igg.theta[3],-1/gamma[3])

prec.V[1:5,1:5]~dwish(G[,],g)
bar.beta[1:5]~dmnorm(beta.prior[],prec.V[,])
}
```

A.2 DMPT: Marginal Posterior Densities Dynamic Effects Parameters

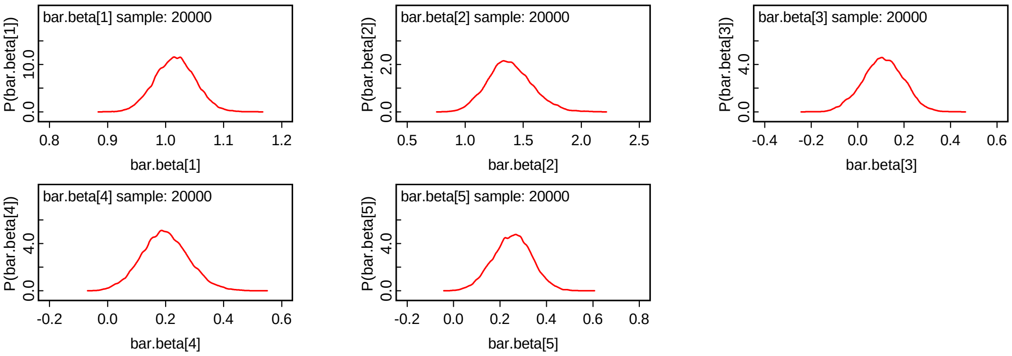

Figure A.1: DMPT/Apparel Marginal Posterior Densities for Vector $\bar{\beta}$ after 50,000 Iterations per Chain (2 Chains, 150K Burn-In)

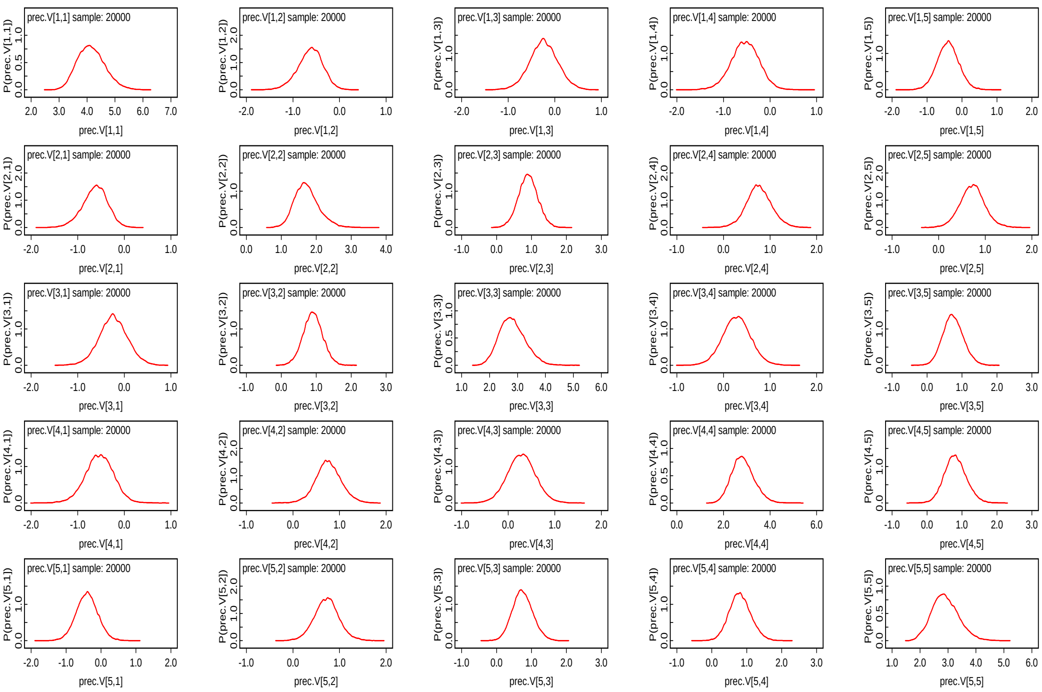

Figure A.2: DMPT/Apparel Marginal Posterior Densities for Precision Matrix V after 50,000 Iterations per Chain (2 Chains, 150K Burn-In)

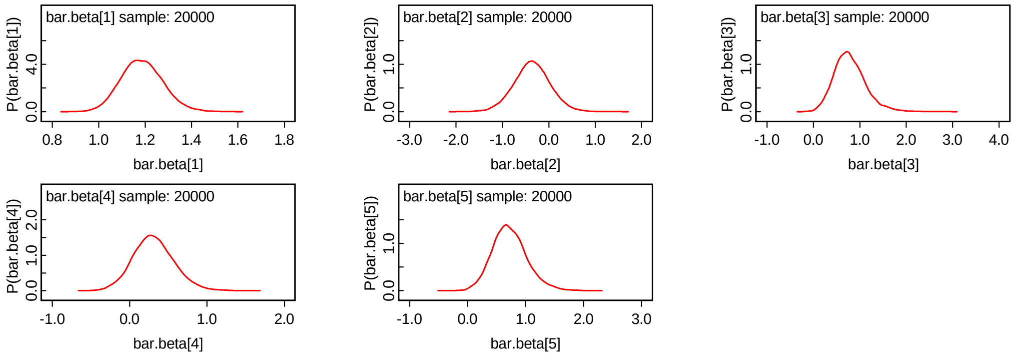

Figure A.3: DMPT/CDNOW Marginal Posterior Densities for Vector $\bar{\beta}$ after 50,000 Iterations per Chain (2 Chains, 150K Burn-In)

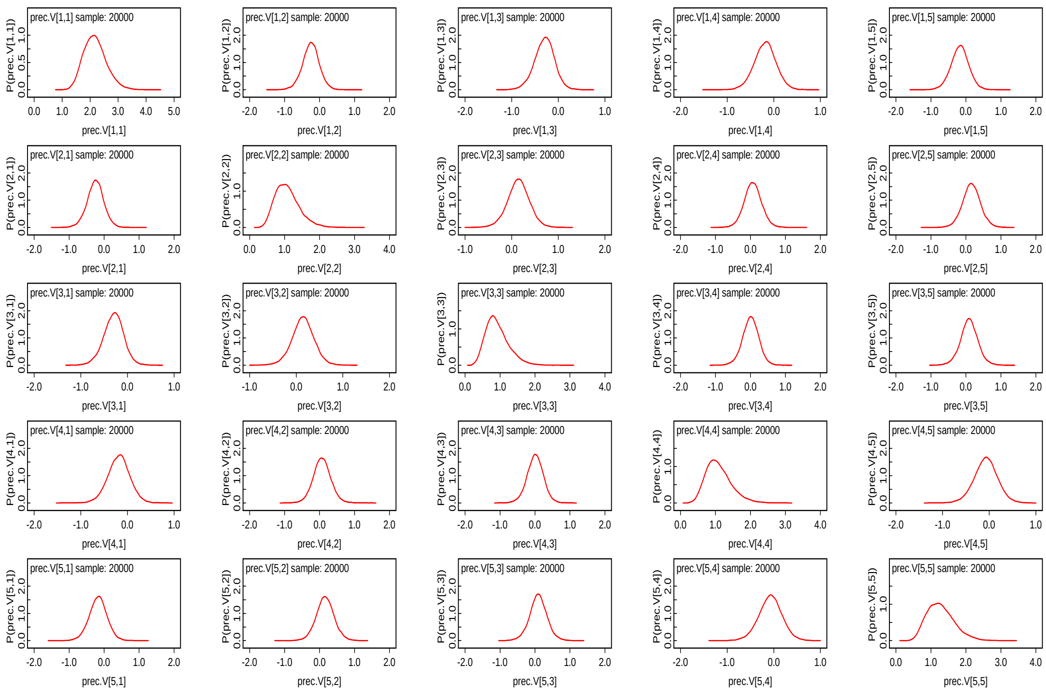

Figure A.4: DMPT/CDNOW Marginal Posterior Densities for Precision Matrix V after 50,000 Iterations per Chain (2 Chains, 150K Burn-In)

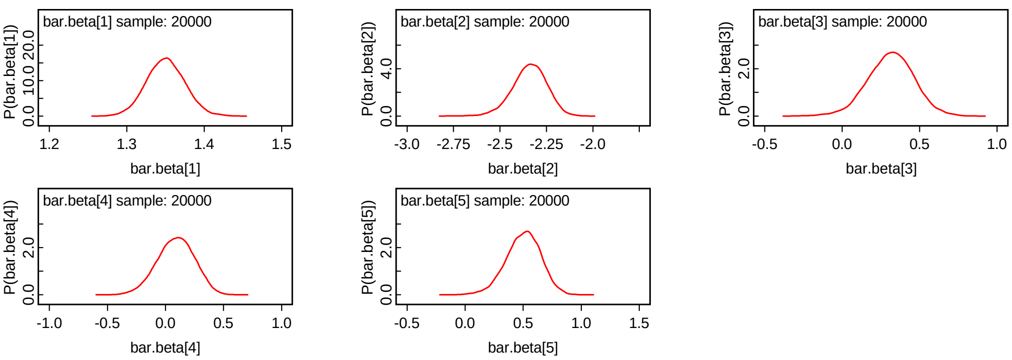

Figure A.5: DMPT/DIY Marginal Posterior Densities for Vector $\bar{\beta}$ after 50,000 Iterations per Chain (2 Chains, 350K Burn-In)

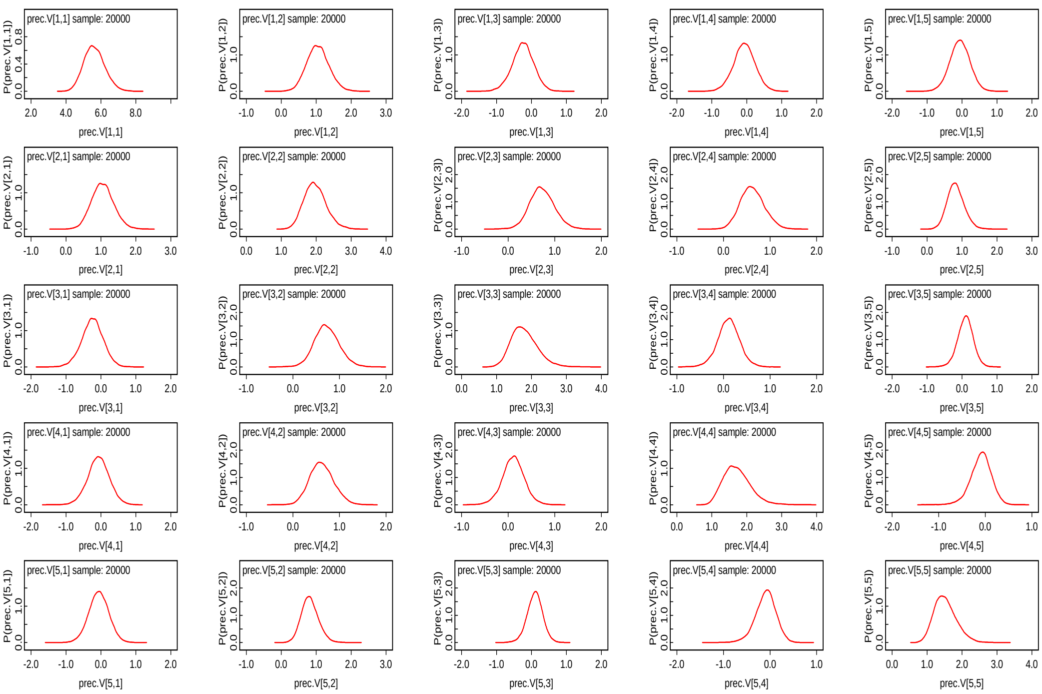

Figure A.6: DMPT/DIY Marginal Posterior Densities for Precision Matrix V after 50,000 Iterations per Chain (2 Chains, 350K Burn-In)

Appendix B

HSM: Implementation, Program Code and Additional Figures

B.1 Implementation Details and JAGS Model Code

The source code to specify the MCMC simulator in the *JAGS* modeling language for the Hierarchical Bayesian Seasonal Effects Model is shown in Source Code Clipping B.1. The code contains the hierarchical Poisson covariate gamma mixture, the hierarchical seasonal effects submodel and all prior definitions. *JAGS* stands for "Just another Gibbs sampler" and is a platform for the MCMC simulation of Baysian models (Plummer 2003). The modeling language is a dialect of the WinBUGS syntax.

The latinized names of the symbols are used in the source code, e.g., α becomes *alpha*. Subscripts are separated by a point, e.g., τ_β becomes *tau.beta*. The variable N denotes the number of individual customers in the dataset and K is the number of seasonal components. The link function $k[j]$ links each period in time j to a seasonal component $k[j]$.

I used "hierarchical centering" of the parameter set $\hat{s}_k$=*s.adj* around zero (line 21), so that s_k reflect the (approximate) seasonal percentage change compared to average purchases levels. Instead of drawing the parameter set *s.adj* directly from a zero-mean distribution the model is over-parameterizing by a floating parameter *mu.s* (line 18). Then *s.adj* is drawn from a normal with mean *mu.s* and precision *tau.s* (lines 19-22). After centering *s.adj* the resulting vector *s* corresponds to the original specification (compare Gelman and Hill 2007, p. 420). This method substantially improves convergence speed of the MC simulation as the sampler traverses the relevant parameter space of this multiplicative model more efficiently (Gelfand, Sahu, and Carlin 1995; Roberts and Sahu 1997; Schliep and Hoeting 2015). For the same reason, the parameters *beta.adj* are first drawn form a normal distribution with mean zero and precision one (line 33) and then transformed to *beta* by multiplying with $1/\sqrt{\tau_\beta}$ and centering at one (line 34).

The parameters *r* and *alpha* are transformations of *mu.lambda* and *sigma.lambda*. The parameter *mu.lambda* corresponds to the location and *sigma.lambda* corresponds to the scale of the gamma(r, α) distribution. This decorrelates r and α and improves effective sample size.

The model can be used to infer only the model parameters from observed data or to simultaneously estimate the model parameters and generate individual forecasts of purchase frequencies. Let T denote the number of observed time periods and let F be the number of forecast periods. Then, the parameter TF (line 24) can be set to T to infer only the model parameters. In order for JAGS to generate forecasts one can set TF to $T+F$ and organize the matrix $x[i,j]$ in such a way that $x[i,j] = NA$ for $j > T$. In this case the model yields parameter estimates and simultaneously generates forecasts of purchase frequencies in the forecast horizon $T+1...T+F$ for each customer *i*.

Listing B.1: The HSM Source Code in JAGS Modeling Language

```
model{

# N       = No. of customers
# K       = No. of seasonal components
# TF      = No. of observed + forcast time periods
# k[j]    = Link time period j to component k[j]
# x[i,j]  = Observed purchase frequencies

tau.s          ~ dgamma(0.001,0.001)
tau.beta       ~ dgamma(0.001,0.001)
mu.lambda      ~ dgamma(0.001,0.001)
sigma.lambda ~ dgamma(0.001,0.001)

# Parameter transform to decorrelate r and alpha
r             <- pow(mu.lambda,2)/sigma.lambda
alpha         <- mu.lambda/sigma.lambda

mu.s          ~ dnorm(0,0.001)
for (m in 1:K) {
  s.adj[m] ~ dnorm(mu.s, tau.s)
  s[m]     <- s.adj[m]-mean(s.adj[])     # 0 - Center
  }

for (i in 1:N) {
  for (j in 1:TF) {
    slambda[i,j] <- lambda[i]*exp(beta[i]*s[k[j]])
    x[i,j]         ~ dpois(slambda[i,j])
    }

  lambda[i]    ~ dgamma(r, alpha)

  # Implies normal(1, tau.beta)
  beta.adj[i] ~ dnorm(0, 1)
  beta[i]     <- beta.adj[i]*(1/sqrt(tau.beta))+1
  }
}
```

B.2 HSM: Autocorrelation and Trace Plots

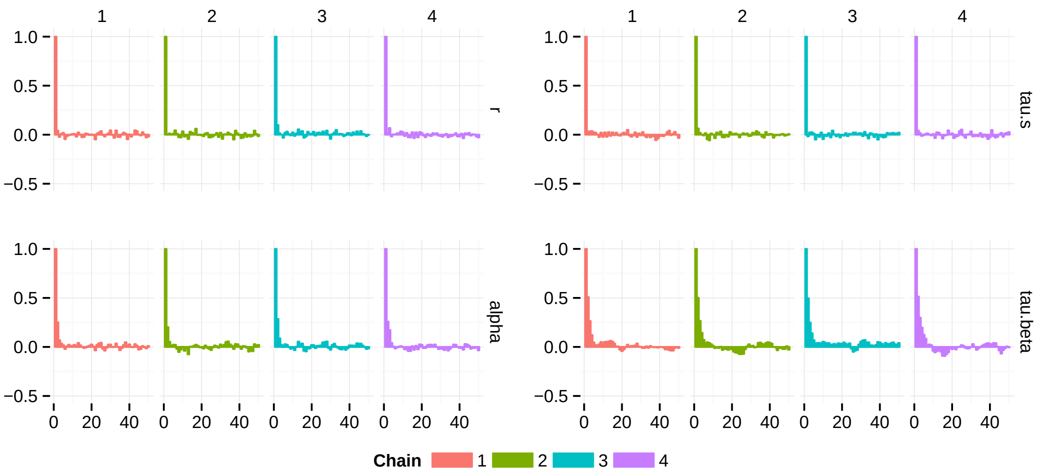

Figure B.1: HSM/Apparel Autocorrelation for r, α, τ_s, and τ_β over 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

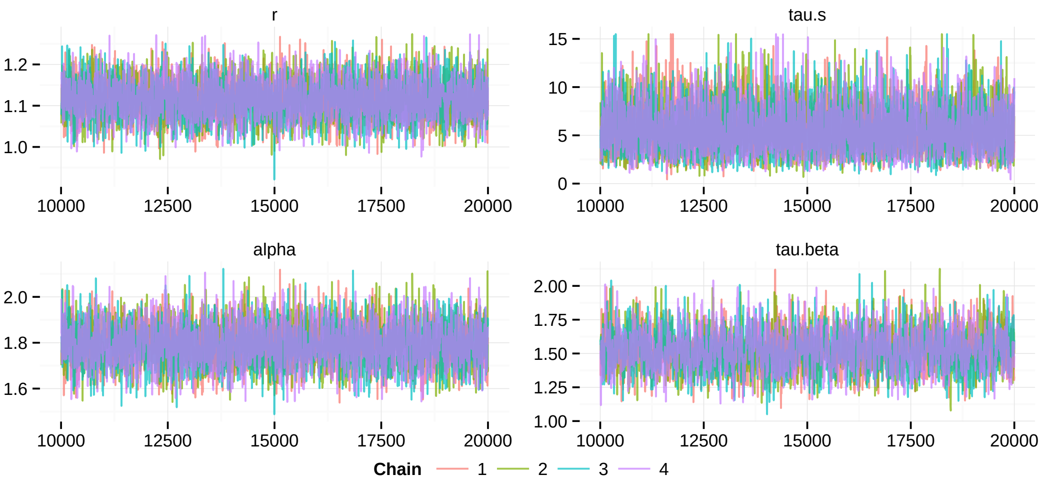

Figure B.2: HSM/Apparel Trace Plots for r, α, τ_s, and τ_β over 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

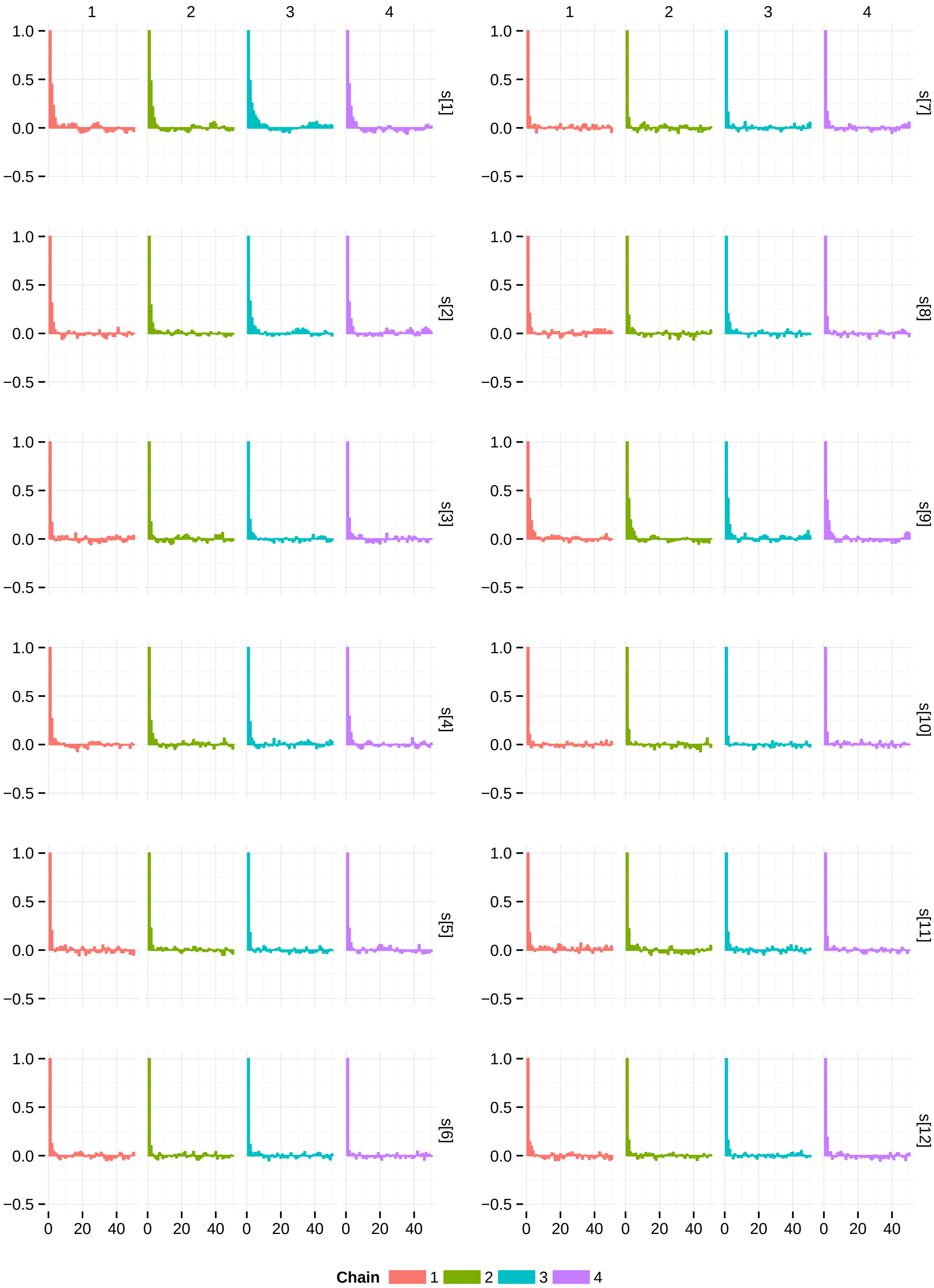

Figure B.3: HSM/Apparel Autocorrelation Seasonal Components s_k over 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

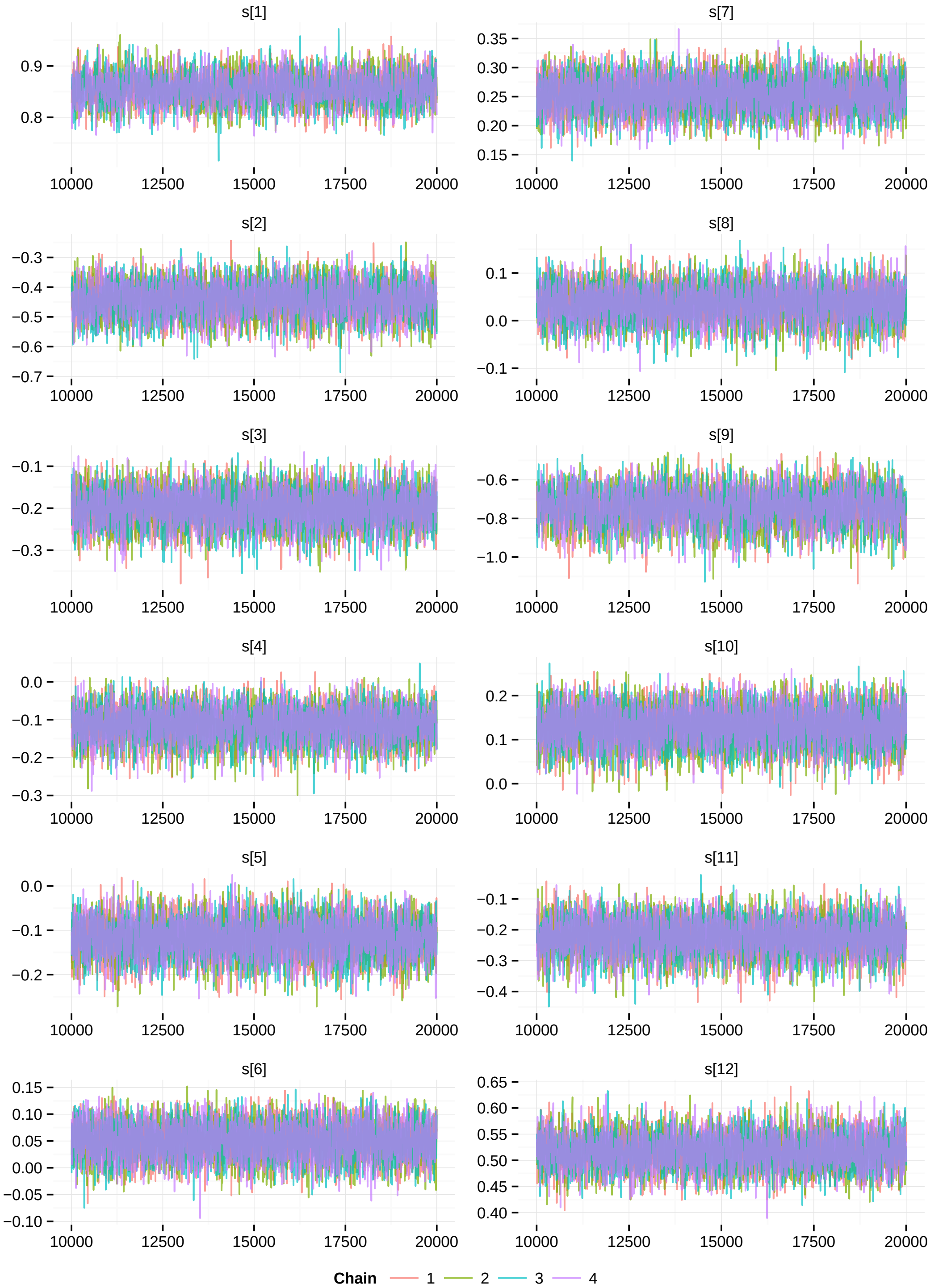

Figure B.4: HSM/Apparel Trace Plots for Seasonal Components s_k over 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

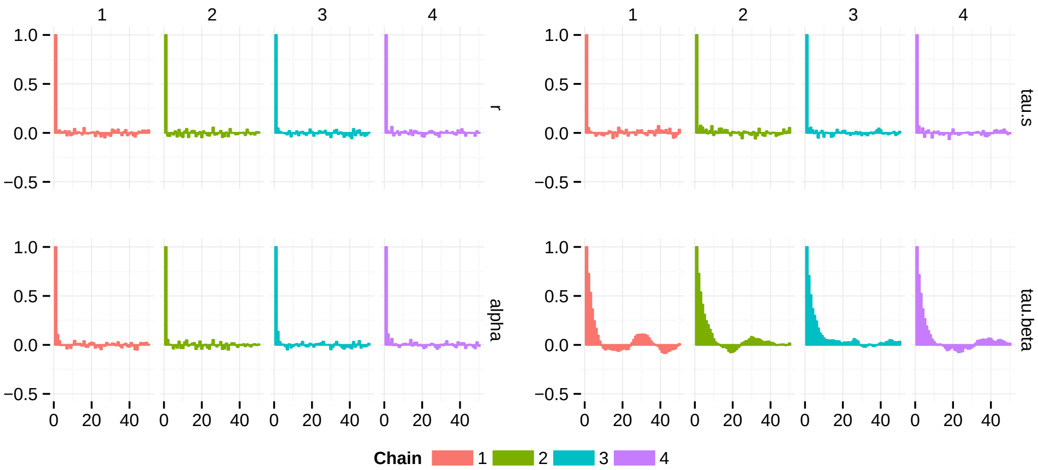

Figure B.5: HSM/DIY Autocorrelation for r, α, τ_s, and τ_β over 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

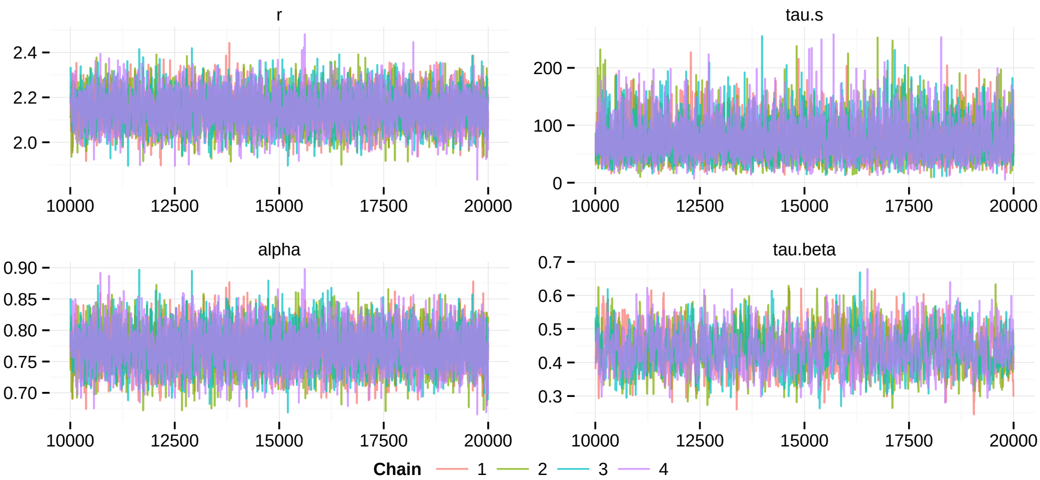

Figure B.6: HSM/DIY Trace Plots for r, α, τ_s, and τ_β over 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

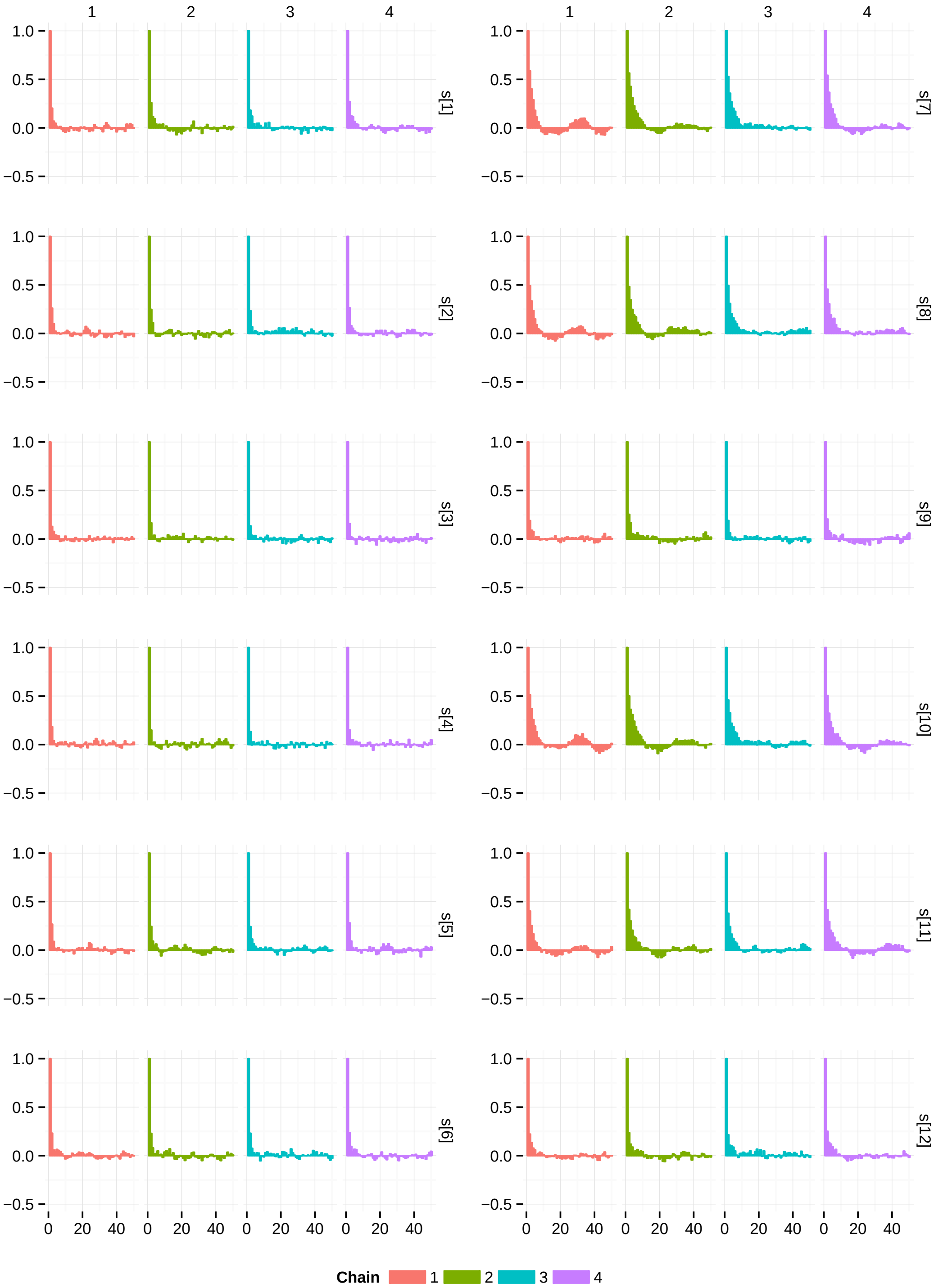

Figure B.7: HSM/DIY Autocorrelation for Seasonal Components s_k over 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

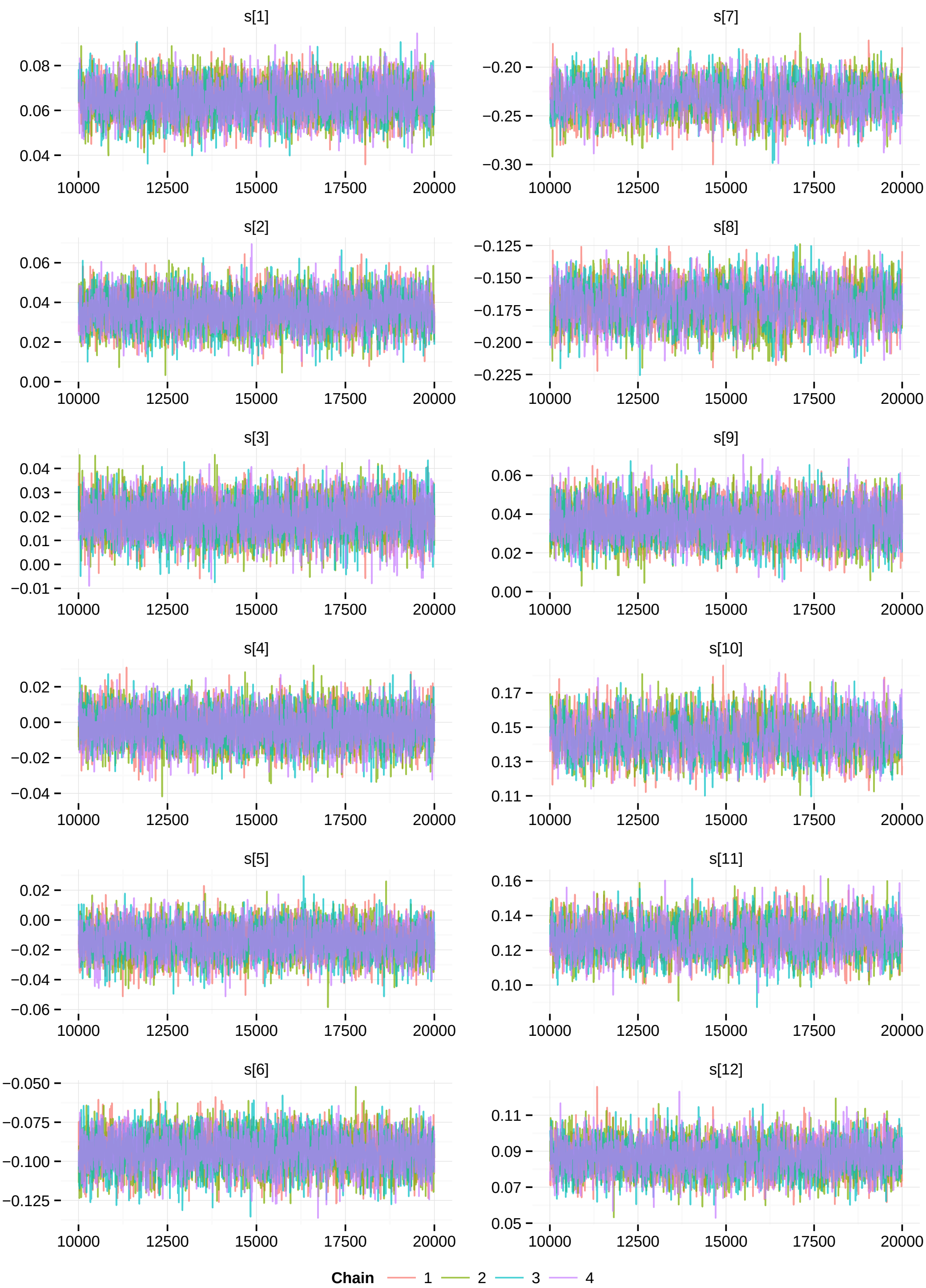

Figure B.8: HSM/DIY Trace Plots for Seasonal Components s_k over 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

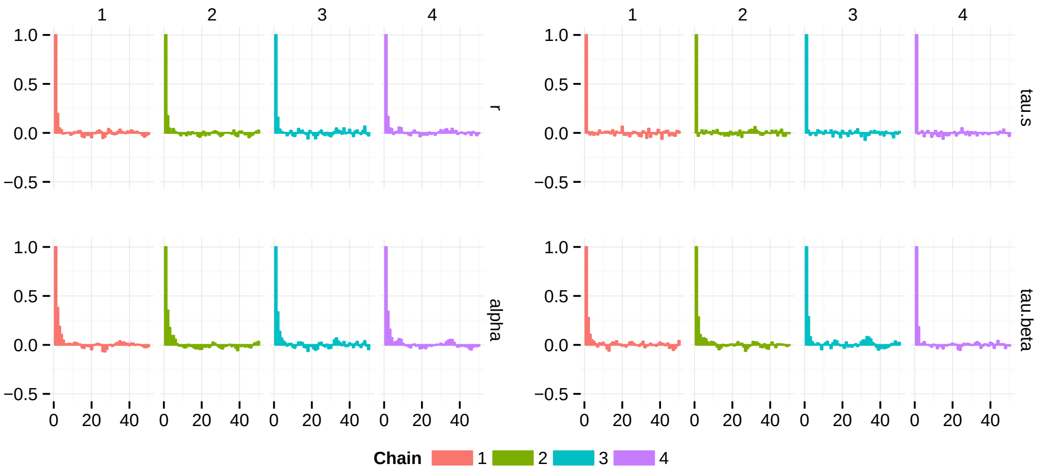

Figure B.9: HSM/CDNOW Autocorrelation for r, α, τ_s, and τ_β over 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

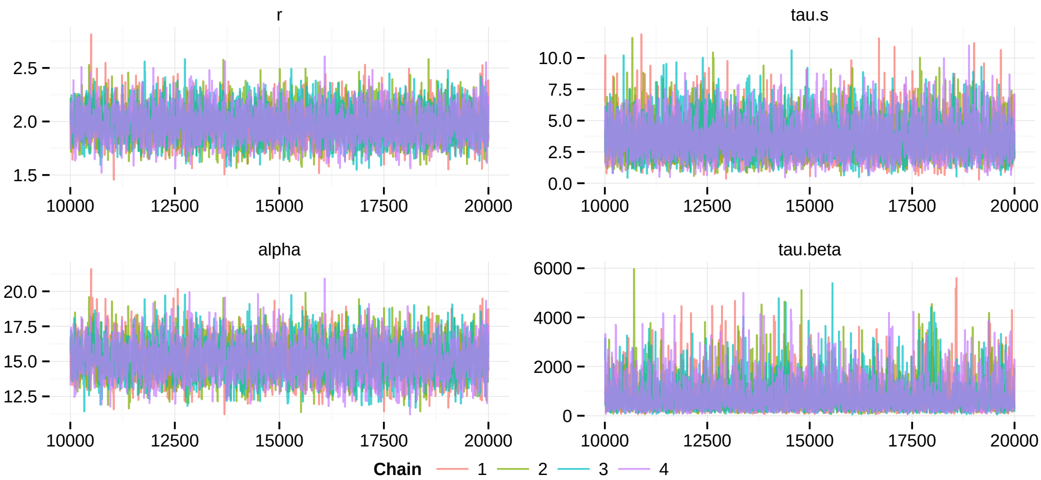

Figure B.10: HSM/CDNOW Trace Plots for r, α, τ_s, and τ_β over 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

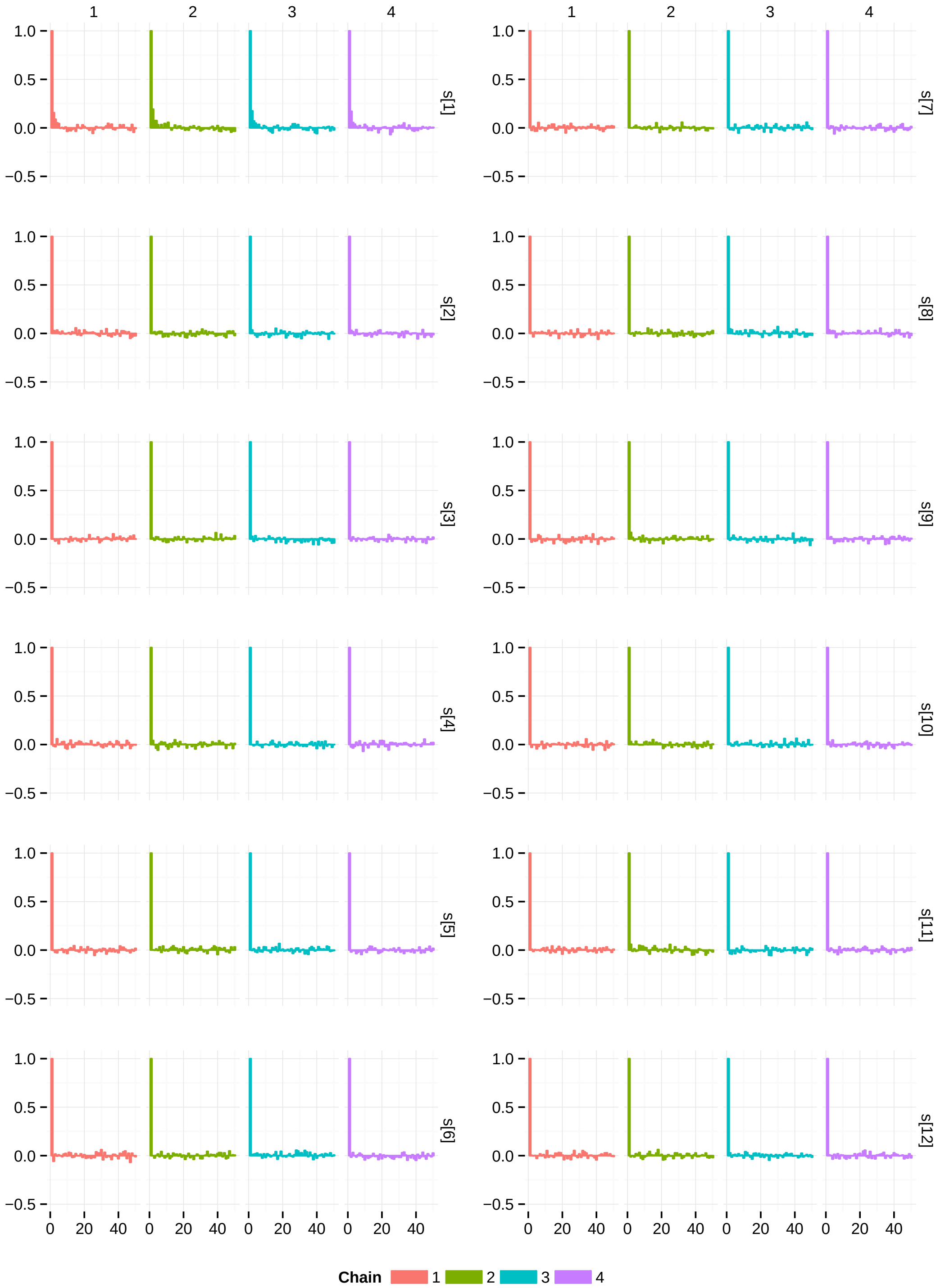

Figure B.11: HSM/CDNOW Autocorrelation for Seasonal Components s_k over 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

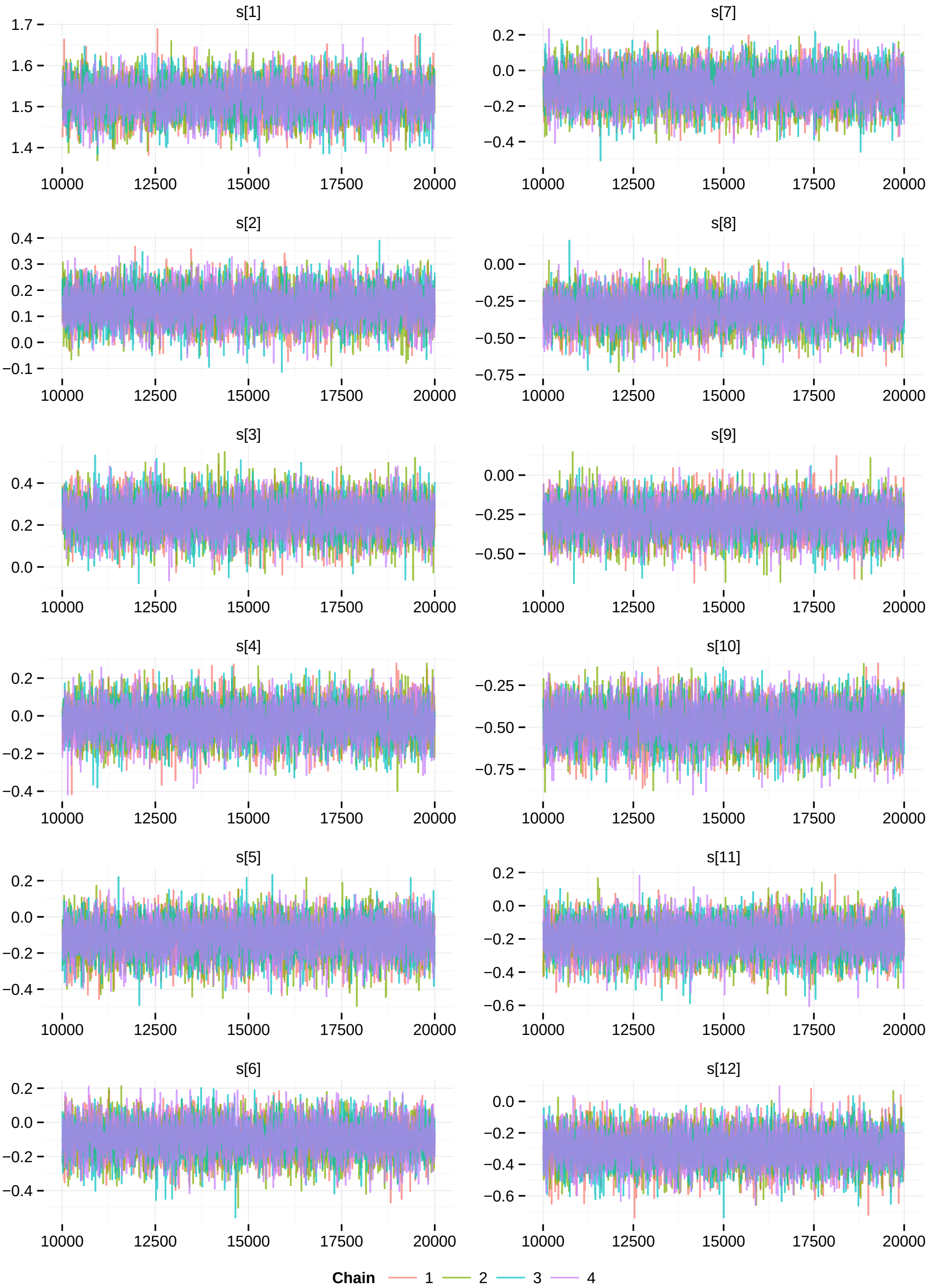

Figure B.12: HSM/CDNOW Trace Plots for Seasonal Components s_k over 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

Appendix C

HSMDO: Implementation, Program Code and Additional Figures

C.1 Implementation Details and STAN Model Code

The source code to specify the HMC simulator in the *STAN* modeling language for the Hierarchical Bayesian Seasonal Effects Model with Drop-Out is shown in Source Code Clipping C.1. The code contains the customer lifetime model, the hierarchical Poisson covariate gamma mixture, the hierarchical seasonal effects submodel, and all prior definitions. *STAN* is a probabilistic programming language implementing full Bayesian statistical inference in a Hamilton Monte Carlo Framework with the no-U-turn sampler (Carpenter et al. 2015; Stan 2015).

Latinized names instead of symbols are used in the source code, e.g., α becomes *alpha*. Subscripts are separated by an underscore, e.g., τ_β becomes *tau_beta*. The variable N denotes the number of individual customers in the dataset. K is the number of seasonal components, T is the number of observed time periods, F is the number of forecast periods, and $t[i]$ denotes the time period with the most recent purchase for customer i. The link function $k[j]$ links each period in time j to a seasonal component $k[j]$.

I use "hierarchical centering" of the parameter set $\hat{s}_k$=*s_adj* around zero (lines 38, 43, 54, and 55), so that s_k reflect the (approximate) seasonal percentage change compared to average purchases levels. Also, the parameter vector *beta* is a transformation of *beta_adj* (lines 39, 42, 56 and 57). The elements of *beta.adj* are first drawn from a normal distribution with mean zero and standard deviation one and then transformed to *beta* by multiplying with σ_β and centering at one (compare Gelman and Hill 2007, p. 420). Note that *STAN* parameterizes the normal distribution with standard deviation instead of precision. Therefore, I transform *tau_s* and *tau_beta* to their respective standard deviations *sigma_s* and *sigma_beta* in the transformed parameters block (lines 38 and 39).

The parameters a and b for the beta distributions are re-parameterized (lines 35 and 36) in terms of a mean parameter ϕ_μ with a uniform (between 0 and 1) prior and a total count parameter ϕ_c with a weakly informative Pareto prior (compare Gelman et al. 2003, p. 128). This decorrelates a and b and improves effective sample size.

The variables *X_total*, *si_xi*, *si_beta* and *l_sum* are intermediate quantities to calculate the model log-likelihood. The full sample log-likelihood, the log of Equation (8.19, p. 133), is calculated in lines 50-74. This encapsulates the computation of the individual log-likelihood (lines 60-73), which is the log of Equation (8.17, p. 133).

The generated quantities block (lines 76-127) is used to calculate the purchase frequency forecasts, $P(Zero_F)$, and $P(alive)$. The vectors *xp_si_beta*, *xp_tau*, *xp_p*, *xz_tau*, and *xpp_si_beta* are intermediate quantities. $P(alive)$ is calculated according to Equation (8.25, p. 135). The forecast of purchase levels uses Equation (8.26, p. 136). Finally, $P(Zero_F)$ is computed using Equation (8.41, p. 142).

Listing C.1: The HSMDO Source Code in STAN Modeling Language

```
data {
  int<lower=1> N;           // No. of customers
  int<lower=1> K;           // No. of seasonal components
  int<lower=1> T;           // No. of observed time periods
  int<lower=1> F;           // No. of forecast time periods
  int<lower=0> k[T+F];      // Link j to component k[j]
  int<lower=0> x[N,T];      // Observed purchase frequencies
  int<lower=0> t[N];        // Recency
  }
transformed data {
  int<lower=0> X_total[N];  // Total purchases per customer
  for (i in 1:N){
    X_total[i]<-sum(x[i]);
     }
}
parameters {
  vector<lower=0>[N] lambda;
  vector[N] beta_adj;
  vector[K] s_adj;
  real<lower=0>    r;
  real<lower=0>    alpha;
  real<lower=0>    tau_s;
  real<lower=0>    tau_beta;
  real<lower=0.1> phi_c;
  real<lower=0, upper=1>    phi_mu;
}
transformed parameters {
  vector[N] beta;
  vector[K] s;
  real<lower=0> sigma_s;
  real<lower=0> sigma_beta;
  real<lower=0> a;
  real<lower=0> b;
```

```
  a<- phi_c*phi_mu;
  b<- phi_c*(1-phi_mu);

  sigma_s<-1/sqrt(tau_s);     // Uses sd instead of precision
  sigma_beta<-1/sqrt(tau_beta);

  // implies normal(1,sigma_beta)
  beta<-(beta_adj-mean(beta_adj))*sigma_beta+1; // 1-center
  s<-s_adj-mean(s_adj);                         // 0-center
  }
model {
  vector[T]    si_xi;
  vector[T]    si_beta;
  vector[T+1]  l_sum;

  r          ~ gamma(0.001,0.001);
  alpha      ~ gamma(0.001,0.001);
  phi_c      ~ pareto(0.1,1.5);
  lambda     ~ gamma(r,alpha);         // Vector operation
  tau_s      ~ gamma(0.001,0.001);
  s_adj      ~ normal(0, sigma_s);     // Vector operation
  tau_beta   ~ gamma(0.001,0.001);
  beta_adj   ~ normal(0, 1);           // Vector operation

  for (i in 1:N) {
    for (j in 1:T) {
      si_xi[j]  <- x[i,j] * s[k[j]];
      si_beta[j]<- exp(beta[i] * s[k[j]]);
      }
    for (tau in t[i]:T) {
      l_sum[tau]<- lbeta(a + 1, b + tau - 1)
                   - lambda[i] * sum(head(si_beta,tau));
      }
    l_sum[T+1]<- lbeta(a, b + T) - lambda[i] * sum(si_beta);
```

```
    // Model individual Log-Likelihood
    increment_log_prob(multiply_log(X_total[i],lambda[i])
                          + beta[i] * sum(si_xi) - lbeta(a,b)
                          + log_sum_exp(tail(l_sum,T-t[i]+2)));
  }
}
generated quantities {
  matrix<lower=0>[N,F] f;   // Forecasts T+1...T+F
  vector[N] PZF;            // P(Zero_F)
  vector[N] PA;             // P(Alive)
  vector[N] p;

  vector[T] xp_si_beta;
  vector[T+1] xp_tau;
  vector[T+1] xp_p;
  vector[F+2] xz_tau;
  vector[F] xpp_si_beta;

  for (i in 1:N)  {
    for (j in 1:T) {
      xp_si_beta[j] <- exp(beta[i]*s[k[j]]);
      }
    for (j in t[i]:T) {
      xp_tau[j]<- lambda[i] *
        (sum(xp_si_beta) - sum(head(xp_si_beta,j))) +
        lbeta(a + 1, b + j - 1) - lbeta (a, b + T);
      xp_p[j]  <- lambda[i] *
        (sum(xp_si_beta) - sum(head(xp_si_beta,j))) +
        lbeta(a + 2, b + j - 1) - lbeta (a, b + T);
      }
  xp_tau[T+1] <- 0;
  xp_p[T+1]   <- lbeta(a+1, b + T) - lbeta(a, b + T);

  p[i] <- exp(log_sum_exp(tail(xp_p,T-t[i]+2))
```

```
                 -log_sum_exp(tail(xp_tau,T-t[i]+2)));
  PA[i]<- 1/(exp(log_sum_exp(tail(xp_tau,T-t[i]+2))));

  for (j in 1:F) {
    xpp_si_beta[j]<-exp(beta[i]*s[k[T+j]]);

    // Forecasts time-period T+j
    f[i,j]<- exp((log(lambda[i]) + beta[i] * s[k[j+T]] +
                  lbeta(a,b+j-1) - lbeta(a,b)) -
                  log_sum_exp(tail(xp_tau,T-t[i]+2)));
    }
  for (j in 1:F) {
    xz_tau[j]<- -lambda[i] * sum(head(xpp_si_beta,j)) +
                 lbeta(a + 1, b +T+ j - 1) - lbeta(a, b+T);
    }
  xz_tau[5]<- -lambda[i] * sum(xpp_si_beta) +
               lbeta(a, b + T+ F) - lbeta(a, b+T);
  xz_tau[6]<- log_sum_exp(segment(xp_tau,t[i],T-t[i]+1));

  // P(ZERO_F)[i]
  PZF[i]<-exp(log_sum_exp(xz_tau) -
               log_sum_exp(tail(xp_tau,T-t[i]+2)));
  }
}
```

C.2 HSMDO: Autocorrelation and Trace Plots

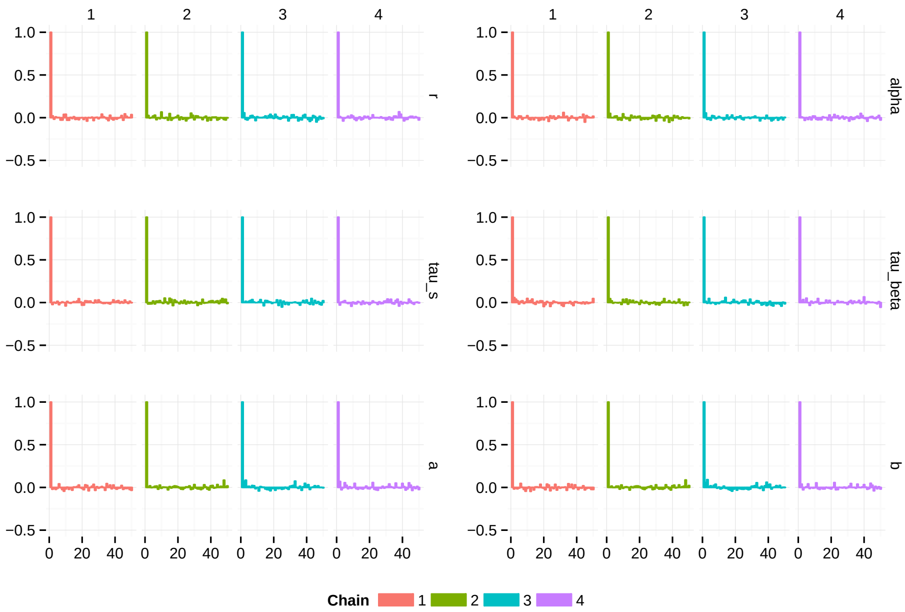

Figure C.1: Apparel Autocorrelation for a, b, r, α, τ_s, and τ_β

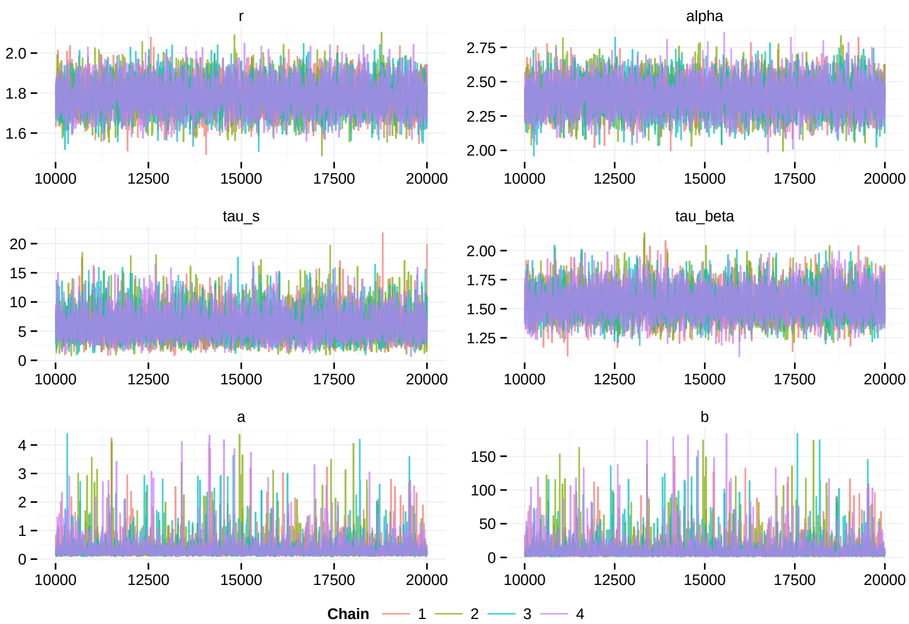

Figure C.2: Apparel Trace Plots for a, b, r, α, τ_s, and τ_β

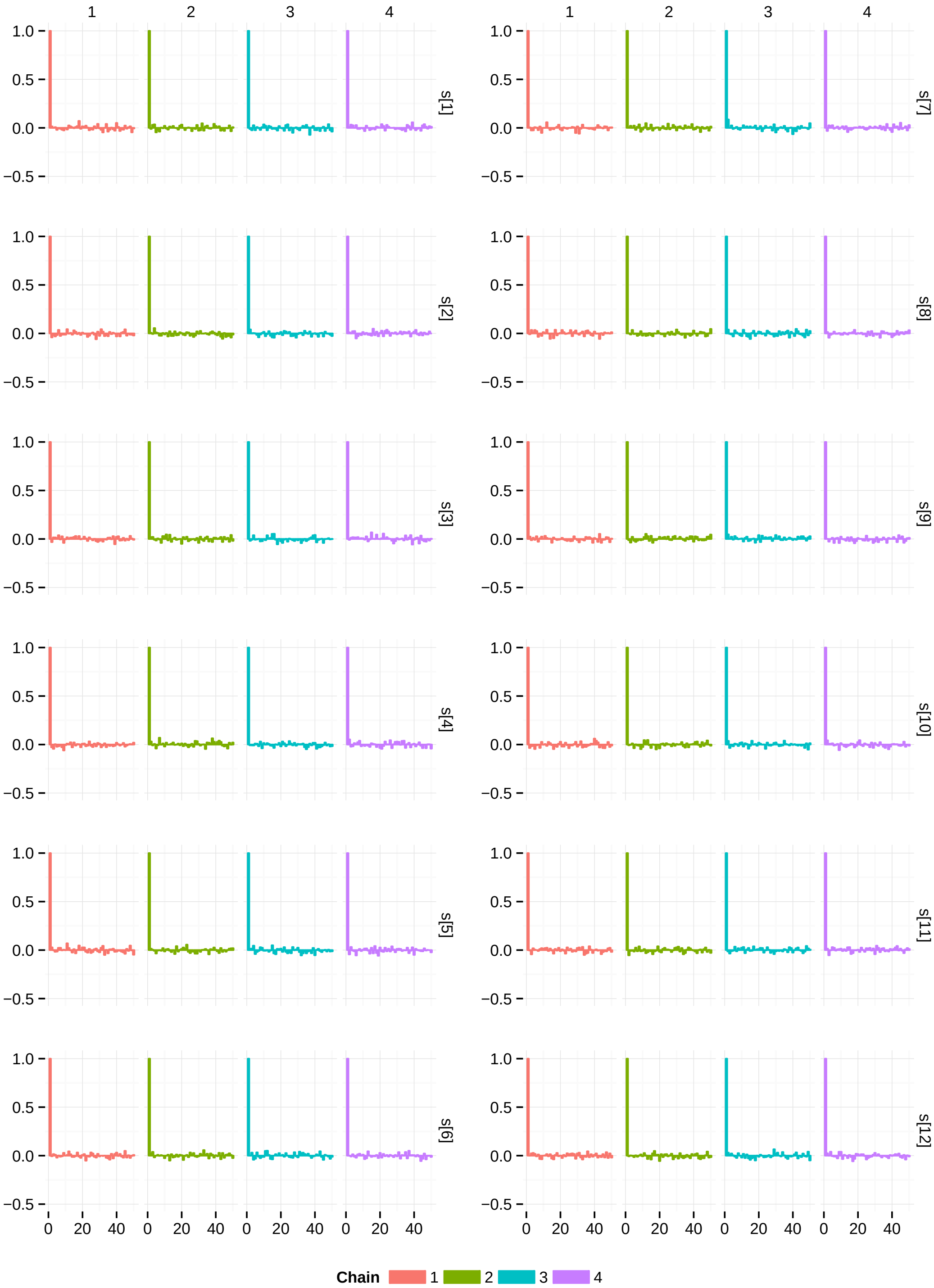

Figure C.3: HSMDO/Apparel Autocorrelation Seasonal Components s_k over 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

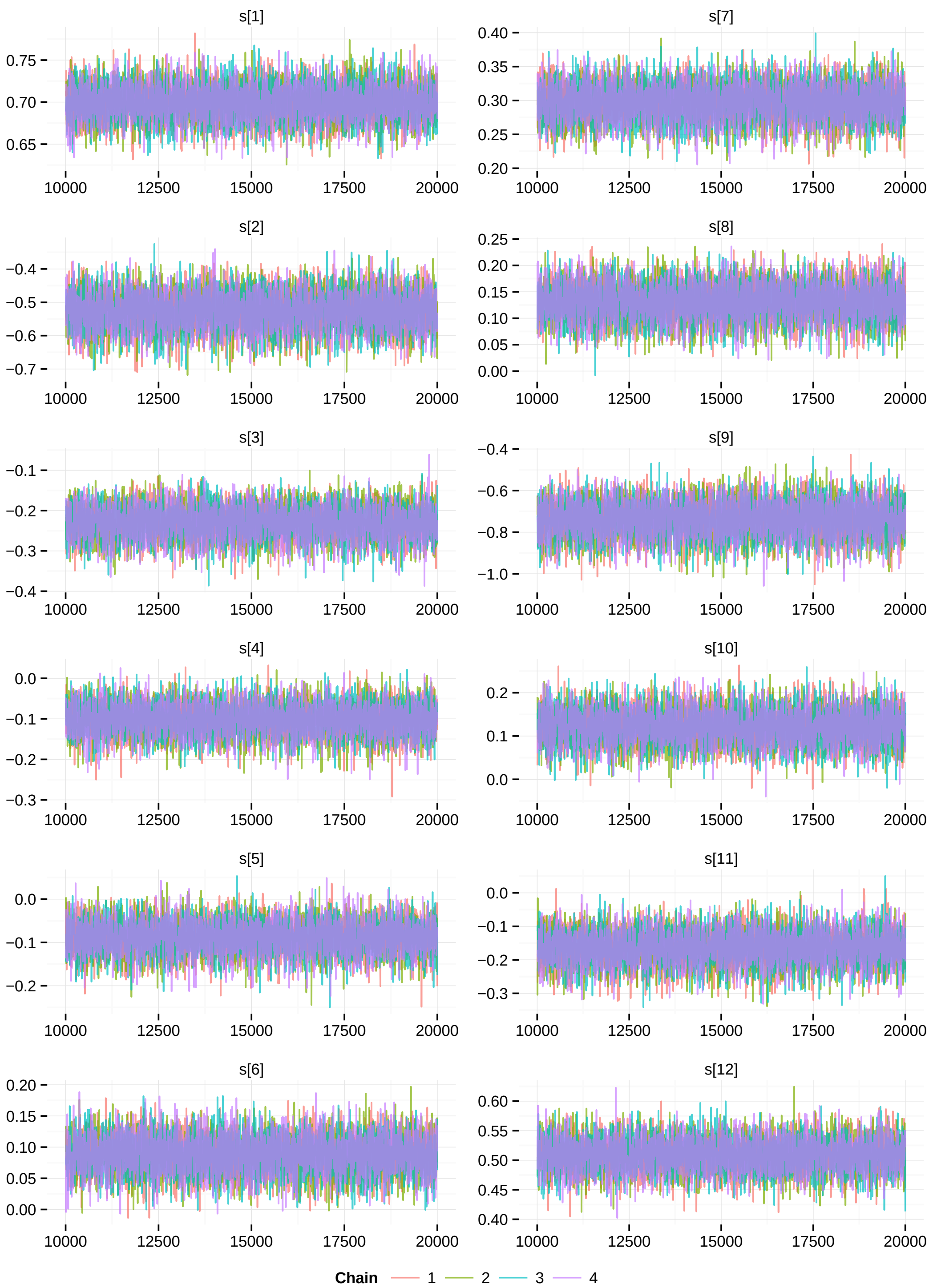

Figure C.4: HSMDO/Apparel Trace Plots for Seasonal Components s_k over 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

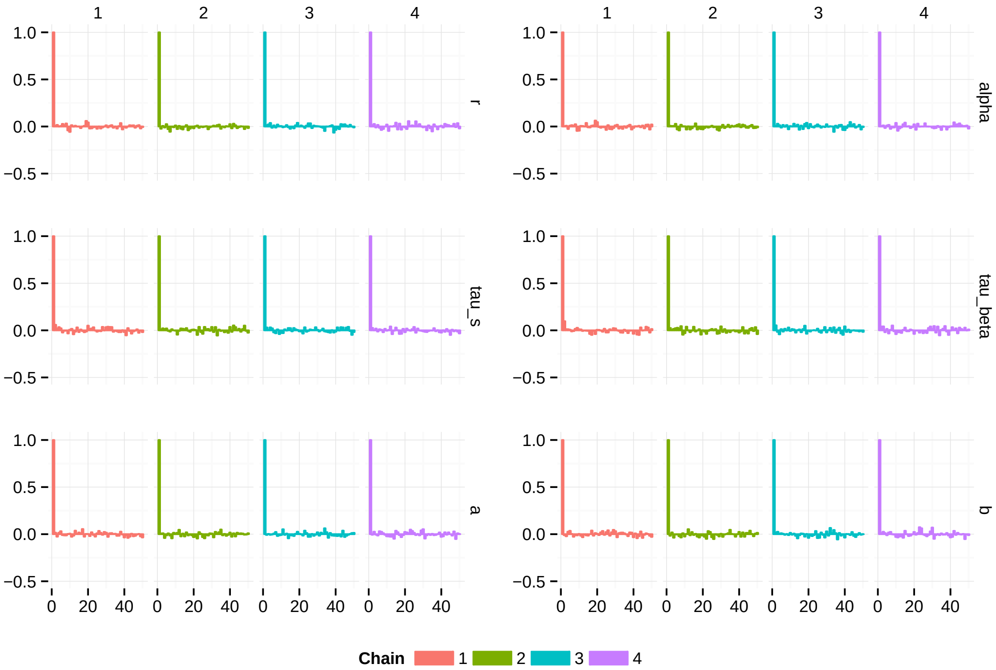

Figure C.5: HSMDO/DIY Autocorrelation for a, b, r, α, τ_s, and τ_β

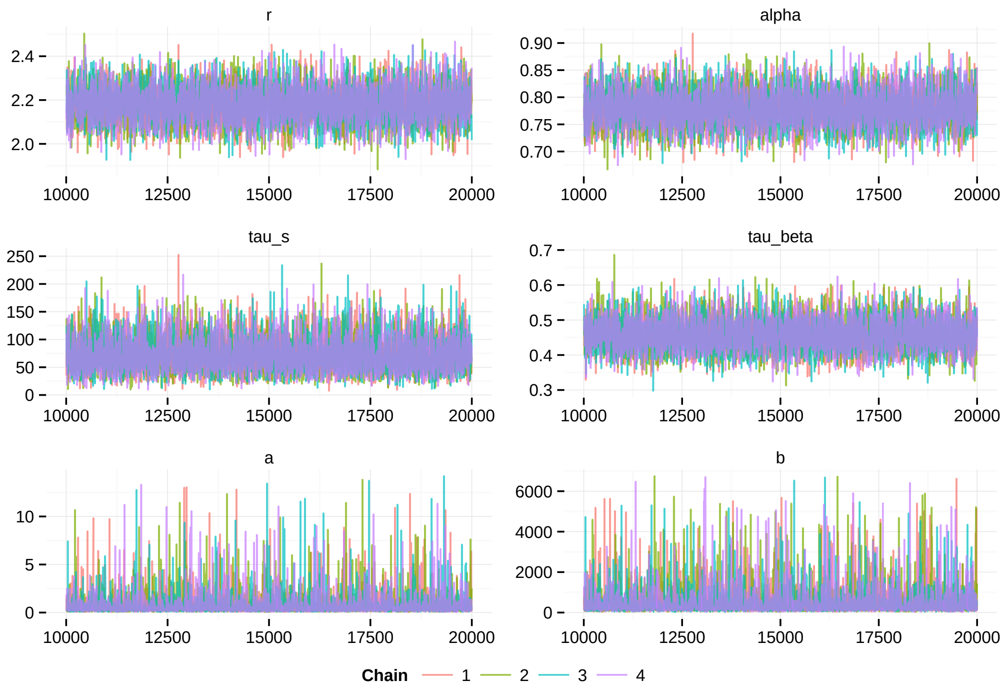

Figure C.6: HSMDO/DIY Trace Plots for a, b, r, α, τ_s, and τ_β

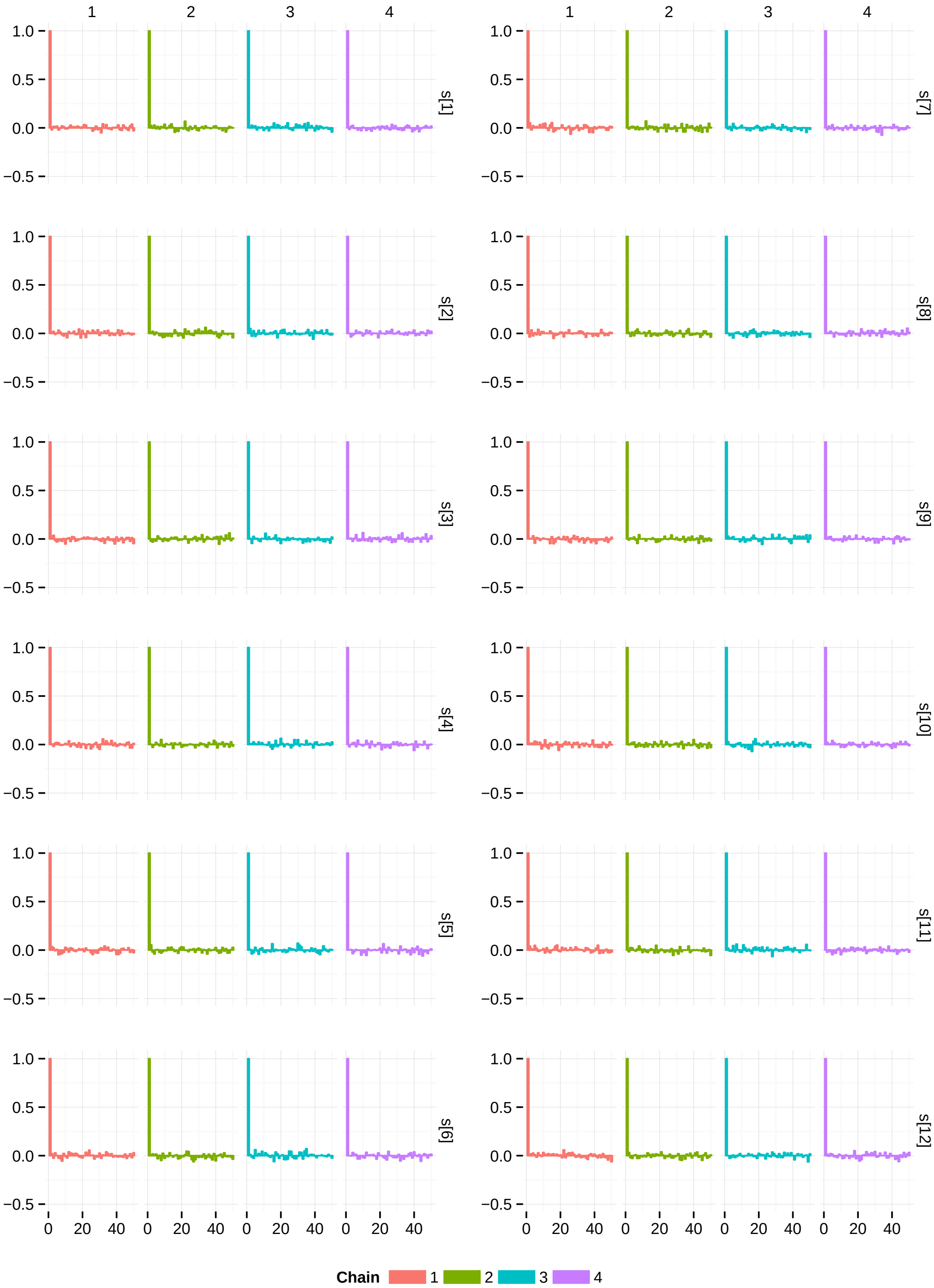

Figure C.7: HSMDO/DIY Autocorrelation Seasonal Components s_k over 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

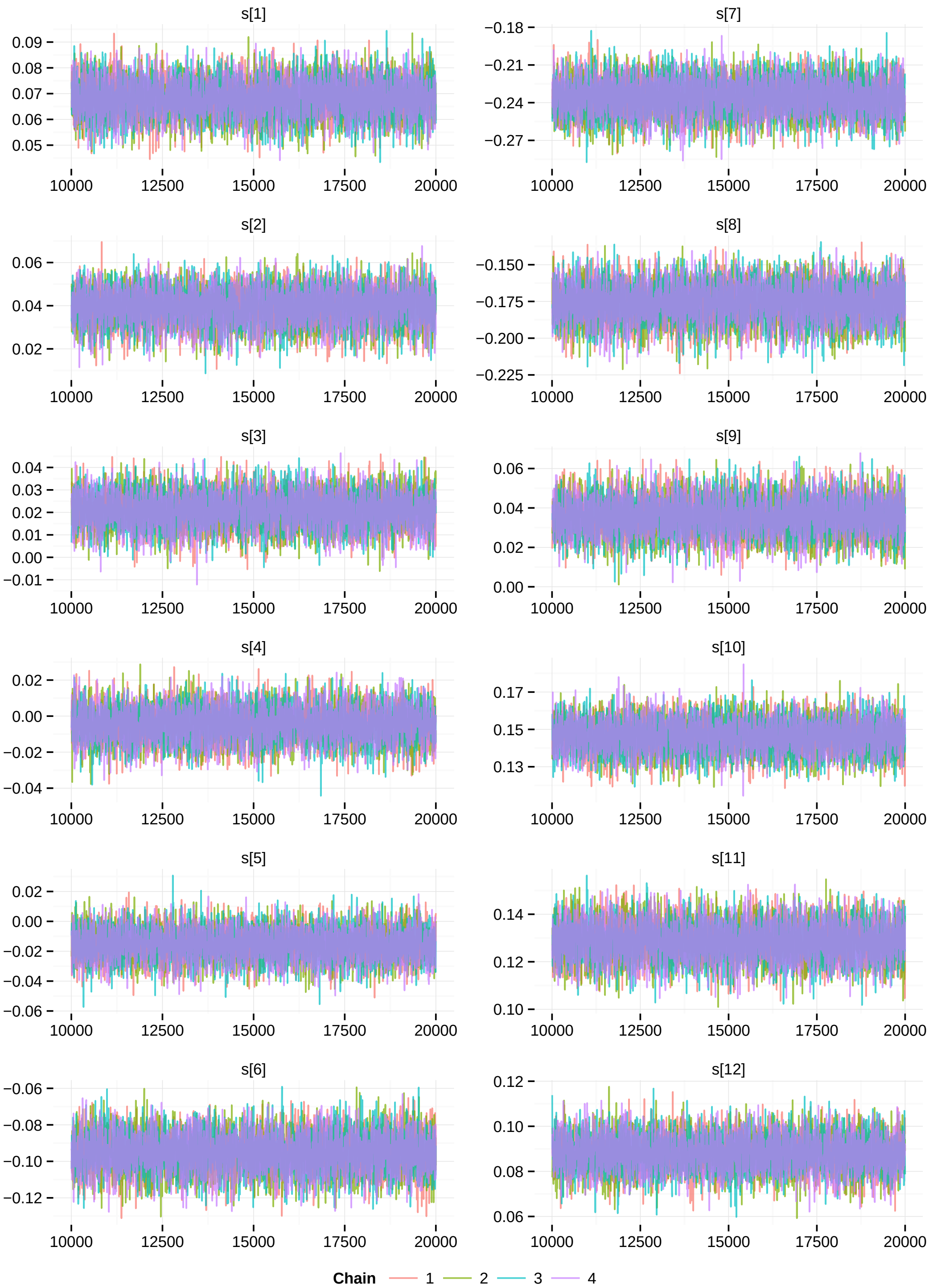

Figure C.8: HSMDO/DIY Trace Plots for Seasonal Components s_k over 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

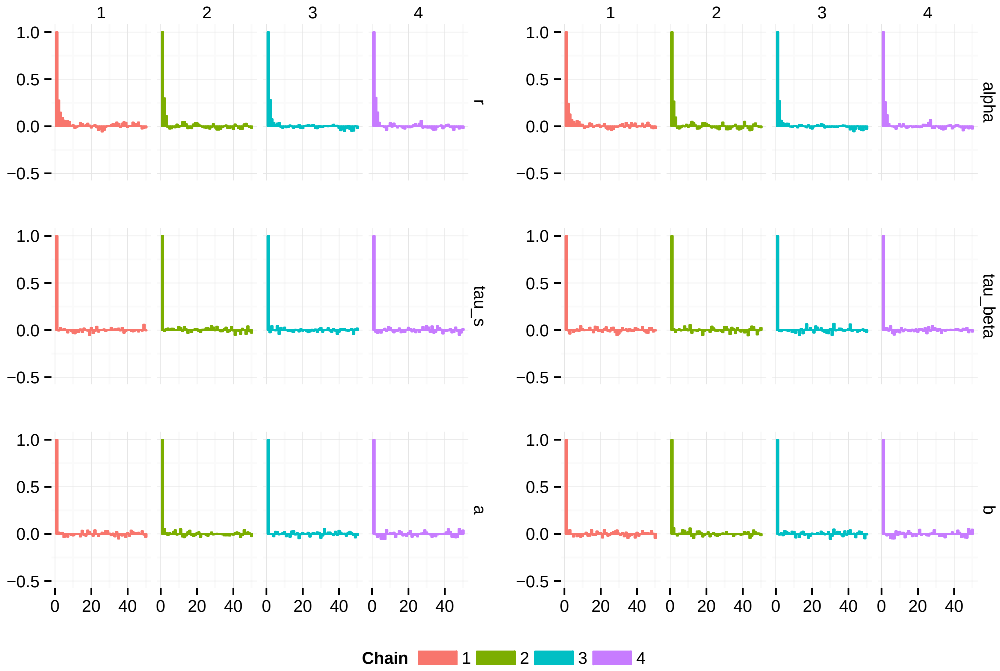

Figure C.9: HSMDO/CDNOW Autocorrelation for a, b, r, α, τ_s, and τ_β

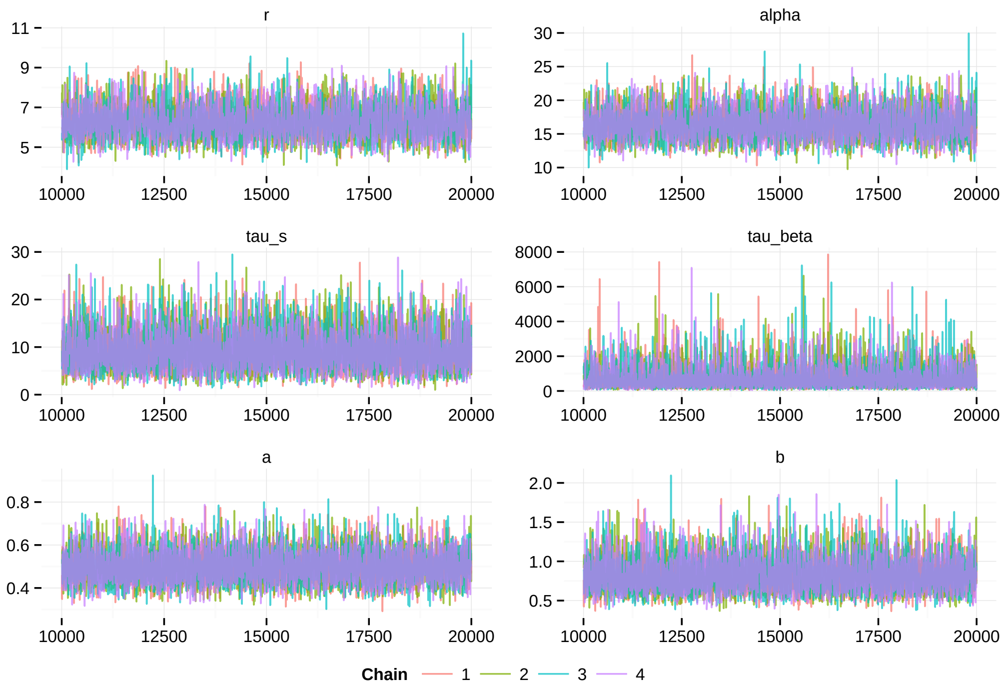

Figure C.10: HSMDO/CDNOW Trace Plots for a, b, r, α, τ_s, and τ_β

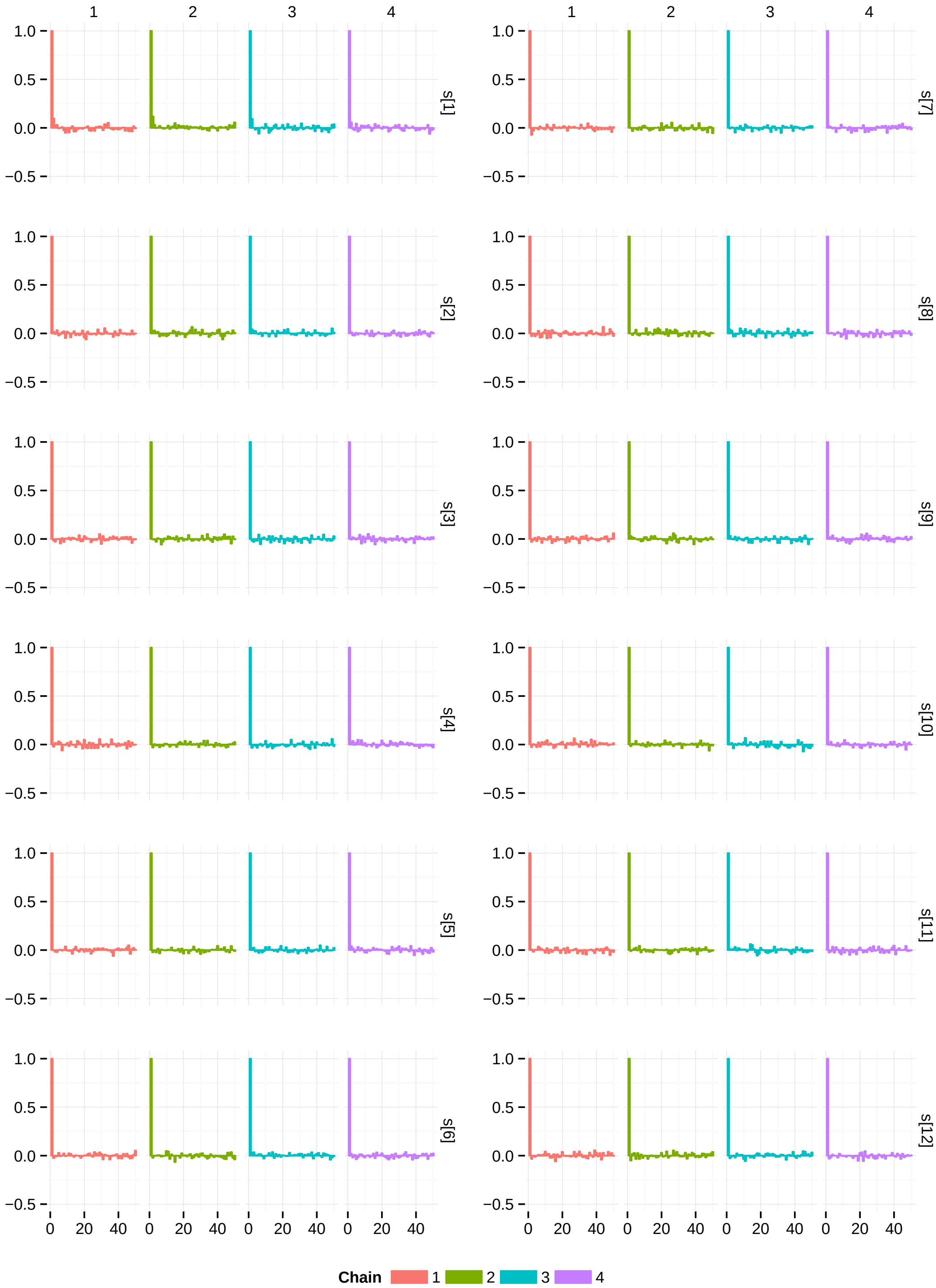

Figure C.11: HSMDO/CDNOW Autocorrelation Seasonal Components s_k over 10,000 Iterations per Chain (4 Chains, 10K Burn-In)

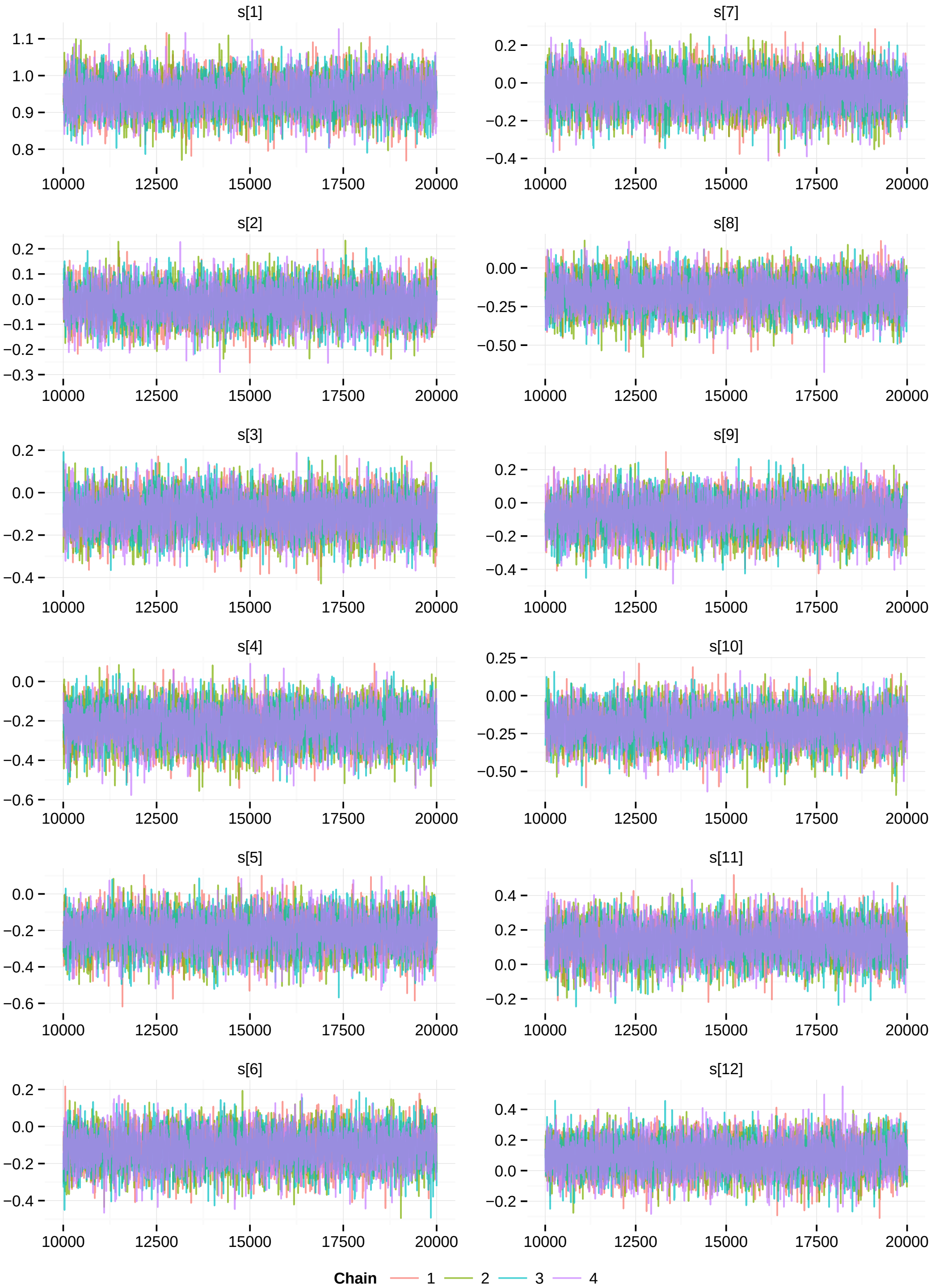

Figure C.12: HSMDO/CDNOW Trace Plots for Seasonal Components s_k over 10,000 Iterations per Chain (4 Chains, 10K Burn-In)