Bayesian Nonparametrics for Marketing Response Models

by

Longxiu Tian

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Business Administration and Scientific Computing)
in the University of Michigan
2019

Doctoral Committee:

Professor Fred M. Feinberg, Chair
Associate Professor Elizabeth E. Bruch
Professor Peter J. Lenk
Assistant Professor Eric M. Schwartz

Longxiu Tian

longxiu@umich.edu

ORCID iD: 0000-0001-6257-7583

In memory of my grandfather Tian Zhidao (田志道), my first teacher.

Acknowledgements

This dissertation would not have been possible without the support of my dissertation committee members, whose guidance and support go far beyond this work. The countless chats, sit downs, emails, and brainstorming sessions that would fundamentally shaped this dissertation – and indeed the course of my PhD – seemed so innocuous at the time. It was Elizabeth who opened my eyes to the vast potential of online dating platforms for marketing and sociological research, and perhaps even more importantly, to the idea that modern methods and contexts can be opportunities to disentangle long-standing questions. Simply put, essay one is a direct fruitation of having Elizabeth as one of my mentors. I am also deeply indebted to Eric, whose far-reaching foresight I've always been able to rely on over the years as I sought to chart out my research direction and focus. Here, Michael Braun deserves mention for not only connecting me to Eric but also for turning my interest towards marketing in the first place when I took his brilliant course on customer-base analysis at MIT. I have also had the good fortune of having Peter Lenk on my committee, a pioneering expert in the the theory and application of Bayesian nonparametrics. To my advisor Fred, I came to Michigan expecting to learn from you about marketing research, choice modeling, and Bayesian inference; of these I've learnt so much from you, as well as a few added lessons I couldn't help but to absorb, like how to conduct oneself professionally and to always treat others decently.

Lastly, to my parents Wei and Jun, there are no words to express the entirety of my gratitude for the enormity of your love and sacrifice. To my wife Sharla, thank you for sharing my laughter and shouldering my worries, as if they were one and the same. At the time of this writing, I'm remain in serious consideration your YouTube channel idea, "Bayesian with Bae".

Table of Contents

List of Figures

List of Tables

List of Appendices

Abstract

This dissertation consists of three essays organized around Bayesian statistical and nonparametric methods to infer latent variables either unavailable or in-principle unobservable, yet nonignorable for robust data-driven marketing decision-making. This work examines applied contexts across ecommerce and online services using large-scale customer-level data from experimental and observational settings.

Its development is motivated by the need to address the intersection of (1) marketing response models to tractably handle the high-dimensionality and volume of customer-level data sources faced by firms today, while (2) enabling robust inference on decision variables in the presence of nonignorable missingness, e.g., those arising from selection, truncation, and censoring. As these missingness typically take the form of partially observable response outcomes, they are unlikely to be fully overcome by the growth in data size alone, or even be feasibly collected (e.g., A/B tests without the existence of an outside option). Compounding this problem, classical methods for addressing nonignorable missingness (e.g., Heckman, covariate-based matching, parametric data fusion) are often hampered in their effectiveness and tractability by the vast scale of these modern datasets. To this end, this work focuses on the development of econometrically principled and computationally scalable Bayesian techniques for robust model-based inference on contemporary marketing data sources and problems, with specific focus on using Gaussian processes as a nonparametric prior over latent variables to flexibly and parismoniously infer the effect of missing data.

The first essay explores how freemium users consider tradeoffs when faced with menu-based pricing using an orthogonalized pricing experiment. Here a data augmentation framework is fashioned for the purpose of recovering the joint distribution over plan choices *by all users* from the conditional distribution of plan choices *by only upgraders*, the latter inherently availed to us by the experiment. Scaffolding on the first essay's parametric framework for missing data imputation, the second essay proposes a flexible Bayesian *nonparametric* data fusion framework over multiple data sources to infer the long-run subscription renewal outcomes of new users.

The third essay addresses a prevalent form of unscored gaps in consumer credit scores, referred to as *sparsely unscoreable*, whereby a consumer's credit score portfolio contains periods that contemporaneously exhibit a mix of observed and missing scores. To address this problem, a Bayesian nonparametric latent factor model is developed to impute credible intervals for gaps in individual score histories within a portfolio of dynamically and contemporaneously interrelated scores.

1 Introduction

It was one of those characteristically Ann Arbor autumn evenings on the drive home with my wife after dinner that an age-old question came up amidst our usual chit-chat. While I no longer remember the lead-up, I distinctly recall asking her (and perhaps making a mental note to mention it in a piece of writing in the future), "why do you think there is something as opposed to nothing?"

This dissertation has a much more limited scope than this question – rather, it focuses on the development and application of modern computational Bayes to address the need by firms to engage in marketing activities robustly, efficiently, and at scale, over time as well as across wide swaths of customers, both in terms of size and preferences. However, it was her answer to this question that has deeply influenced my research interests and direction since that evening. Her response is central to the cohesion of this work; if you will, the shrinkage prior over my dissertation. What she said remains personally the singularly most succinct summary of the underpinnings of Western philosophy and thinking that I had ever heard to this day. To my question, she responded, "because existence is simply *better* than non-existence."

That is, it is more '*good*'. The philosophical, theological, and indeed, cultural implications of her response are countless. But for our purpose, I propose a simple takeway: the notion of good is fundamental. It is fundamental to how we've chosen to engage with our surroundings, form our societies, and organize our markets. In this dissertation, the notion of good is understood in terms of both as an adjective (i.e., the quality of adding utility) as well as a noun (which taken together with services, constitute *products*). As an adjective, I will assume that it is the axiomatic driver of consumer behavior, of which however cannot be presumed uniform, entirely deterministic, nor stationary over time. As a noun, it is what firms proffer to consumers at a cost, and composed of characteristics that are possessed across the set of all competing alternatives (e.g., price, quality, features, and other covariates) as well as those that are idiosyncratic (e.g., brand- or choice- specific effects), but all of which will be presumed measureable, additive, and thereby, compensatory on

some utility scale. To the likely readers of my dissertation, these are familiar, if not unnecessarily rehashed, classic random utility and choice modelling primitives, whose roots can be traced to the works of Thurstone (1927), Luce (1959), McFadden (1974), Guadagni and Little (1983), among others.

I've nonetheless chosen to make explicit these premises because I wish to highlight that the so-called 'modern' computational machinery covered in the following chapters have actually been developed to solve problems in marketing that one may be surprised to find as having remained outstanding to this day, despite their pervasiveness (i.e., menu pricing, polytomous responses with sample selection, unscored gaps in credit scores) or perspicuity (i.e., fusing CRM databases to A/B tests for longitudinal analysis). This work as well as my original decision to pursue a marketing doctorate are part of the growing interest and advancement in empirical marketing models in recent years, as they've lent themselves well to the increasingly digitized and data-driven relationship between consumers, products, and firms. However, a perusal of Winer and Neslin's *The History of Marketing Science* will show that research into the application of empirical models to understand and manage marketing actions is nothing new, and traces its origin to over a half-century ago when "modelling consumer choice behavior … you were pretty much limited to a few dairy panel datasets." The empirical contexts explored in this dissertation have indeed evolved from these humble beginnings, as now "in the 21st century, we have unlimited data … on e-commerce choice behavior." The implications to this are twofold. First, the methods developed in the following chapters can be directly traced to extant forays into well-known problem domains (e.g., sample selection, data fusion, nonignorable missingness), but have been warranted as classical methods are increasingly hampered in their effectiveness and tractability due to the scale of "unlimited" modern datasets. The three essays will attempt to provide a review of the relevant research and methods onto which they extend, and the limitations they overcome. Second, these methods are designed to shed insight onto new contexts wherefore not possible with existing approaches (e.g., allow for novel substitution patterns, convex menu price optimization, accounting for dynamics in fusing panel datasets, correlated latent variables across time and consumers), but as will be shown, also recover known or expected behavioral and policy phenomena (e.g., IIA violations, choice inertia, the use of short-run elasticities systematically misattribute future revenue across cohorts). In both essays, the reader may find that

the richness of the datasets can engender many more research questions and class of findings than those considered here. However, among the multitude of possible directions, it is around the *applicability* of these models to support data-driven decisions by managers that this research is conceived and organized. That is, the focal objective is the development of marketing response models.

Hanssens et al. (2005) define these as empirical models of how consumers individually and collectively respond to marketing activities. These models take as inputs marketing and environmental variables, the latter of which could consist of characteristics such as customer demographics and their usage patterns, and outputting measurable responses such as preferences, choices, and revenue. As originally conceived, these models are meant to sit at the heart of a decision system whereby objective functions such as those on profit, churn, and firm growth can then be optimized.

Given this definition, two final points on the conceptual framework of this dissertation are worth detailing: the *customer-choice occasion* as the basic unit of analysis, and the role of Bayesian nonparametrics for marketing response models. On the former, consumers' actions are consider individually across discrete choice occasions, as opposed to collectively over a certain time horizon, finite or otherwise. It's worth noting the latter perspective has spawned a range of successful, and indeed scalable, hazard-type models taking advantage of parsimonious cohort-level sufficient statistics (i.e., RFM) to conduct customer-base analysis and predictions (cf. Schmittlein et al. 1987, Fader et al. 2005, Fader et al. 2010). The decision here then to model at the customer-choice occasion level is motivated (1) by the need to conduct optimization over the effect of marketing actions on individual consumers, and (2) that hazard-type models for multinomial choice contexts, as is the empirical setting of essays one and two, lose the computational efficiency of their binary anaologues (i.e., closed-form solutions), resort to restrictive assumptions (e.g., series of binary comparisons), or in many cases, having to incorporate discrete choice components (cf. competing risk models). In the case of essay one, the discrete choice model framework engenders an optimization setup that considers both what is *good* for the firm (i.e., need to maximize profits) with what is *good* for the customer (i.e., the utility of choosing competing alternatives, if any), thus allowing for a rational framework for managerial decision-making based on these findings. In essay two, individual choices are conceptualized as stochastic realizations within the *time-space*

of all possible customer-choice occasions, whose indices are over the set of input features (e.g., pricing, demographics, timing variables such as seasonality, renewal occasion, membership length). This naturally lends to the use of metrics to evaluate the (dis)similarity across choices, both observed and unobserved, including those between disparate datasets, where neighboring data can then be drawn together, or fused, to enable inferences that might have been impossible with any single dataset alone.

At the heart of the data fusion method proposed in essay two is a multidimensional Gaussian process prior (GPP). This Bayesian nonparametric prior can be understood as the similarity metric across observations, as well as the mechanism for the shrinkage, or the *sharing* of information, across datasets. However, it's worth noting that these are not distinct mechanisms of the GPP but rather, one and the same. The use of the GPP for data fusion represents a generalization of extant techniques, such as nearest neighbor algorithms and hierarchical Bayes. In particular, it generalizes data fusion to (parsimoniously) account for different time-scales, and to interchangeably tradeoff among them, a concept which is explored in-depth in the essay. More broadly, the GPP overcomes the linearity assumption common to nearest neighbor algorithms along with the parametric assumptions of hierarchical Bayes, which taken together, serve to alleviate some of the less desirable implications of assuming compensatory utility functions, as established earlier. These advantages of using the nonparametric GPP are certainly not limited to data fusion, but rather what will be covered in the following chapters can be understood in the broader context of robust and flexible inference on marketing response models. As such, while a common inferential theme to the essays is the need to overcome nonignorable missingness – one arising from sample selection, the other truncation over time, yet another the amalgamated effect of inadmissible or incomplete inputs whose functional form is likely highly nonlinear – here, another advantage of nonparametric Bayes arises: scalability. Essay one leverages a parametric imputation technique to recover the missing information on unobserved outcomes. While the Hamiltonian Monte Carlo-based data augmentation strategy is able to overcome the tremendous difficulty of integrating over the high-dimensional joint distributions across latent utilities, such sampling-based treatments limit its applicability to more curated datasets, such as those arising from an experiment, as in our case. In part building on the experience of essay one,

essay two takes advantage of one of the most successful approximate Bayes implementations to date: sparse variational Gaussian process (Titsias 2016). It is presently one of the few known variational approximations with guaranteed asymptotics to the true distribution on a nontrivial model, reducing the notoriously cumbersome $\boldsymbol{O}(n^3)$ estimation of GPPs to $\boldsymbol{O}(nm^2)$, where n in our case is in the order of millions due to the use of a full longitudinal CRM database, whereas in comparison m shall only be in the order of hundreds, enabling a full-scale undertaking difficult for even extant data fusion methods. Essay three leverages hierarchical Gaussian processes to differentially model unobserved carryover whose composition is unlikely to be uniform across customers or time periods, but yet results in a common pathology where consumers cannot be assigned a valid credit score. These contexts is one of many where scalable Bayesian nonparametrics may be of benefit to undertake inference where scalability has thus far hampered robust inference.

Having now established the underlying principles and themes of this research, the remainder of this dissertation is organized as follow: essay one, titled *Optimizing Price Menus for Duration Discounts: A Subscription Selectivity Field Experiment*, is covered in Chapter 2; essay two, titled *Broadening the Horizon: Augmenting One-Shot Field Experiments with Longitudinal Customer Data*, is covered in Chapter 3; and essay three, *Improving Credit Score Forecasts when Data are Sparse: A Dynamic Hierarchical Gaussian Process Model*, is covered in Chapter 4. In these chapters, supporting tables and figures referenced in the body of the essays are provided after the conclusions. Additional supporting details, including detailed derivations, algorithms, and source codes, are provided in the Appendix.

2 Optimizing Price Menus for Duration Discounts: A Subscription Selectivity Field Experiment

2.1 ABSTRACT

Online services typically offer an array of contract durations as a price menu, rewarding longer commitments with lower per-period costs. Key to their success is setting component prices to both encourage potential customers to "select in" overall and to nudge those that do to the best mix of contracts. Gauging inter-plan substitution effects using existing customers' data suffers from low historical frequency of menu pricing updates and component plan prices changing in lockstep. Here, we avail of a formal pricing experiment that orthogonalizes the elevation and steepness of price menus for a major online dating pay site.

While this alleviates collinearity issues intrinsic to the firm's historical pricing policies, it cannot correct for differential customer self-selection, which is a function not only of the absolute menu prices levels, but how they relatively interrelate. We address this via a novel selectivity model that allows for correlated binary selection (purchasing) and multinomial choice (among three available plans), along with numerous individual-level covariates available to the firm at the time of the customer's decision. To perform data augmentation over constrained latent utilities, as well as efficiently recover highly nonlinear parameters and full covariance matrices, the model is estimated using Hamiltonian Monte Carlo.

Parameter estimates and resulting inter-menu price elasticities for a wide range of models suggest that the usual random-utility framework underlying "nonlinear pricing" choices can entail, in addition to diminished within- and out-of-sample fit, discernible artifacts in inferred substitution patterns. The proposed modeling framework allows the measurement of certain anticipated pricing effects (e.g., higher prices discourage purchase overall and choice of any higher-priced plans), but also ones prior models fail to capture: raising the price of the longest-duration contract actually increases subscriptions overall, and "elevating" the price menu (i.e., raising all prices) hurts subscription rates less than raising the medium-duration plan price alone. In

particular, across-the-board pricing increases have a far lower negative impact than standard random-utility models would imply. Finally, optimization of the entire price menu suggests that the firm is setting prices too low overall, particularly for its longest-duration plan, and adjusting the price menu accordingly should lead to roughly 10% greater revenue.

2.2 INTRODUCTION

Firms offering subscription services face a dilemma: they wish to attract a stable base of long-term paying customers, but few are willing to commit before suitably assessing the firm's array of benefits. Various "try before you buy" solutions have thereby arisen to temper the tension between free or teaser-rate trial and contracted revenue commitment. A classic example is the shareware model for software products, offering costless limited functionality, with intermittent reminders to spring for the full version. Such methods can work well when customers pay a one-time fee and "it's yours forever", but less so when a service provider seeks an ongoing revenue stream via fixed-term commitments. Firms thereby wish to optimize the full range of prices offered to incoming customers when first presented with the opportunity to subscribe.

In such situations, as modeled subsequently and operationalized via a field study, individual consumers can self-sort into which service commitment duration, and associated price level, best suits their projection of future needs. Such trade-offs are particularly severe for "apps", where only 5% of total users self-select into purchase, effectively subsidizing the rest (The App Association 2018). Although the marginal cost of an additional user for an online service is typically negligible, the firm's fixed costs are not: to be sustainably profitable, the proportion of paying customers must be suitably high, or the price they each pay sufficiently large, to support the entire enterprise.

Such pricing contract strategies have seen rapid adoption among online services and mobile businesses (Niculescu and Wu 2014), but their workings remain "poorly understood" (Kumar 2014), especially in terms of how consumers make the choice to select into paying, and then subsequently how long to commit for. To entice consumers to "lock in" for a longer period – and thus provide a guaranteed revenue stream – firms typically offer lower per-period costs for consumers committing to

lengthier commitments. The practice spans widely disparate categories: Dronemobile, the smartphone-car tethering service, offers 1-year ($59.99), 2-year ($99.99), and 3-year ($129.99) upfront commitment terms, as does the venerated Chicago Manual of Style ($39, $70, and $99, respectively); World of Warcraft offers modestly declining monthly subscription rates of $14.99 (1 month), $13.99 (3) and $12.99 (6), as does the Washington Post at $9.99 (1) and $8.33 (12) and numerous other journalistic outlets.

Given the ubiquity of such *menu-based duration pricing* for subscription services, firms have surprisingly little controlled experimental data, let alone dedicated modeling capability, on which to calibrate their understanding the inter-plan trade-offs made by consumers. The most obvious of these trade-offs is whether to self-select into paying for a subscription at all; but this depends on the price levels of the various contract offerings. Ideally, the firm seeks to understand how the entire price menu affects consumers' overall decision to subscribe and, conditional on doing so, which subscription terms represent the best balance between risk (longer commitment) and reward (lower per-period usage). The user pondering a lengthy commitment is kin to the supermarket club shopper pondering a super-sized sack of sugar, but without the ability to store, trade, or increase usage rate, a topic we return to later in discussing prior literature on volume discounts and price discrimination.

Compounding firms' measurement problem is that they only rarely alter prices – keeping them stable sometimes for years – and, when they do, typically do so for all service contract lengths in lockstep. Indeed, such is the case for the focal firm in our forthcoming price experiment. Historical subscription data enabling price menu assessment thereby suffer two serious deficits: low temporal pricing variation overall, and high cross-plan price trend correlation. This central problem was underscored by Levitt, List, Neckermann & Nelson (2016), whose study "differs fundamentally from the existing empirical literature because we actually had the power to change prices and did so in a randomized field experiment in which quantity discounts were varied", highlighting that "In contrast to almost all previous studies, we were able to observe not only market outcomes but also the individual actions" of consumers. Keeping plan prices stable for long periods then altering them contemporaneously may be wise for business, but bedevils measurement of cross-plan price substitution effects. Ideally, one would wish to decouple the mean (i.e., typical or representative)

price of a firm's subscription options from the variation across them, in order to extract the cleanest signal of how the price menu acts in concert: to convert subscribers *overall*; and to *redistribute* those paying consumers across available plans.

Here, we overcome these limitations in two ways. First, we examine initial conversion and contract choice using data collected from a subscription field experiment conducted on a menu-priced online dating site; customers are presented a menu consisting of 1-, 3-, and 6- month subscription options, along with corresponding prices, whose levels are manipulated using a factorial design to orthogonalize their *absolute* vs. *relative* values. Second, resulting data are analyzed via a novel methodology assessing the correlation between the latent propensity of whether to subscribe and latent utilities for *which* plan to choose. The modeling framework extends the recent literature on selection effects, e.g., via control functions (Petrin & Train 2010) and multinomial selectivity (Feinberg, Salisbury & Ying 2016), and compares favorably with classic methods used in the price discrimination literature (as detailed later). The model is estimated via Hamiltonian Monte Carlo (HMC), which allows the efficient recovery of even highly nonlinear parameters and full covariance matrices, with source code made available in the extensible probabilistic programming language, Stan.

The chapter is organized as follows. We first review relevant literature in nonlinear pricing, subscription service models, price discrimination, and selectivity effects. We then introduce the setting and particulars of our field experiment, followed by the specific model used to isolate pricing effects of interest. Model estimation, plan price elasticities, and results are then presented and discussed, followed by joint optimization of the full price menu for two firm-relevant objective functions. We conclude with implications for practice and further research into discrete price menu effects.

2.3 SELECTED LITERATURE

The literature on pricing policies in marketing and economics is vast, and even a summary of the research on online pricing is beyond our purview (see, for example, Rao 2009, or Ratchford 2009). Part of the complexity concerns the nature of contractual services, which (in contrast to tangible products) can be altered or

gradated inexpensively or on the fly, even for a large consumer base (e.g., setting a heterogeneous usage cap, as is common in cellular data contracts). The "space" of potential price and service attribute strategies can thereby be so enormous as to be practically infinite.

The use of field experiments to study promotions and pricing is a common practice by marketers and has received extensive treatment in the marketing literature. For example, in the context of subscription services, Danaher (2002) conducted a long-term experiment on the pricing of wireless phone plans to derive a revenue-maximizing strategy for usage vs. retention, while Anderson & Simester (2004) used a series of field experiments to assess the effects of price promotions on future purchasing by first-time and existing customers. Such experiments are a mainstay in economics (Levitt & List 2009) and provide something of a "gold standard" for the assessment of pricing effects (e.g., Reilly 2006).

2.3.1 An Idealized Experiment

In our experimental design, described more fully in the sequel, new registrants of the focal site were randomly assigned to treatment conditions that vary in the levels comprising the price menu. Our goal is to measure both (1) how the entire roster of the menu's component prices affects users' decision(s) to select into a paid subscription, and (2) how all visitors – not only those who decided to subscribe – trade off across subscription contracts. The key observation is that the first of these decisions alters the subset of customers whose trade-offs can actually be observed, i.e., the second decision. For example, suppose that the firm "elevates" the price menu by increasing all component prices by some percentage. It follows then that visitors with higher price sensitivity will differentially self-select out of subscription entirely, meaning that cross-plan substitution patterns will be determined (stochastically) by customers with higher willingness-to-pay. As such, one might see an overall tamping-down of cross-elasticities among the plans, and obtain biased counterfactuals regarding the effects of altering (relative) prices among the plans themselves. In the case of the well-worn multinomial logit model (Guadagni & Little 1983), all such substitution effects depend on the error terms in the latent utilities: the IIA property would suggest that making the outside good (i.e., the "no choice" option) relatively more attractive would leave the ratios of choice probabilities of the other plans largely unchanged. [We will compare the results of the proposed

model to such an alternative in our empirical application.] In short, econometrically "zeroing out" the latent correlation in utilities for selection (into subscribing) and for individual plans can produce constrained elasticity measures, and thereby distort – or worse, yield misleading conclusions for – key managerial exercises, such as price menu optimizations.

One way to conceptualize eliminating such a selection "bias" (in the sense of Heckman 1979) is to imagine an idealized experiment where price menus were suitably manipulated, but where first-time site visitors were required to choose a plan, perhaps by being given a sufficient initial endowment in an incentive-compatible set-up. In this case, the large proportion of visitors who elect not to subscribe in the real world – roughly 80% in our experiment – would provide cross-plan substitution information, and thereby de-bias the counterfactuals necessary for price menu optimization, since there would be no "selecting out". By postulating a latent utility for the dichotomous "choice vs. no choice" that is itself a function of available covariates, its covariance with the latent utilities of the inter-plan choices allows for the extrapolation to the customers whose plan choices are "missing non-randomly" (Little & Rubin 2002; Zanutto & Bradlow 2006). The model developed here "stochastically imputes" the unavailable latent utilities for the vast majority of customers who do not subscribe, conditional on their own covariates and the data-augmented latent utilities for the customers who do, as is possible in the Bayesian framework (Wachtel & Otter 2013). Specifically, to recover the effect of menu-pricing on all first-time visitors in the field experiment, we develop a model of selectivity correction of multinomial choice outcomes.

2.3.2 Nonlinear Pricing, Quantity Discounts, and Contract Duration

As mentioned earlier, the key trade-off for potential subscribers is that committing to a longer plan – say, 6 months – provides a lower monthly cost, but higher total initial outlay. The literature on nonlinear pricing (Wilson 1993) and consequent price discrimination is marketing and economics is extensive, and the reader is directed to the excellent reviews by Iyengar & Gupta (2009) and Lambrecht et al. (2012). In the former's general definition, "a nonlinear pricing schedule refers to any pricing structure where the total charges payable by customers are not proportional to the quantity of their consumed services". Moreover, in regard specifically to price discrimination, consumer heterogeneity is singled out as the

primary motivator for nonlinear pricing, where such structures "can be thought of as a menu of quantities and corresponding charges" (p. 356), and as such directly applies to our empirical context.

A critical issue in assessing quantity discount mechanisms is evaluating the (cross-) elasticities among options presented to consumers. In many tangible-goods contexts, this is complicated by a lack of perfect substitution among options in a product line (Mussa & Rosen 1978), which can differ in nontrivial ways that precludes their being organized along a unidimensional trade-off spectrum, e.g., the styling of different automobiles. Even when this is possible – say, for larger sizes of otherwise identical packaged goods like branded breakfast cereals – those who opt for larger sizes may be purchasing for an entire family, anticipating an increase in usage rate, or have ample storage capacity. For service contracts, users are by contrast locking themselves into different durations, and instead need to project need for a specific future time period. Regardless of context, models assessing such trade-offs hinge on whether they explicitly incorporate a selection mechanism as well as how flexibly they account for unobservable utility shocks among the selection decision and various plan options. Prior research in the area varies considerably in this regard, to which we next turn our attention.

2.3.3 Selectivity and Intercorrelation in Multinomial Choice

Non-random (self-)selection is a well-documented issue in field data settings, and approaches to correcting the resulting selectivity bias have received extensive attention since Heckman's (1979) pioneering work. Extensive reviews have appeared in cognate disciplines (e.g., Heckman 1990, Winship & Mare 1992) as well as in Marketing proper (e.g., Danaher 2002, Wachtel & Otter 2013). Accounting for selectivity specifically in online marketing is increasingly recognized as critical whenever customers self-select into a "treatment". For example, Lambrecht et al. (2011) analyzed banking decisions for customers with online accounts; Braun & Moe (2013) corrected for selectivity in latent rate of exposure to ads and downstream conversion probabilities; Manchanda, Packard & Pattabhiramaiah (2015) correct for potential differences in unobservables between members and non-members of an online community. In all such cases, consumers were not assigned randomly to conditions, thereby requiring post hoc correction.

While experimental random assignment (e.g., treatment or price conditions, etc.) can alleviate certain selection biases, others can remain if measurements are even partially influenced by participants' decisions, even if the various "conditions" have been carefully designed, orthogonalized, and randomly-assigned (Feinberg, Salisbury & Ying 2016). The literature on price discrimination, quantity discounts, and nonlinear pricing relies on a mixture of experimental and observational data; of analytical, structural, and econometric analysis; and on a wide range of modeling techniques. Here, we focus on the models for consumer choice, specifically in how they accommodate (non-random) selection and latent correlations in utility "shocks" (i.e., errors or disturbances). One method for handling this self-selection is to view it as an explicit "branch" on a decision tree. For example, Train et al. (1987, 1989) pioneered the use of the nested logit model in this regard, to study consumer choice among phone plans, first determining the plan then, conditionally, the level of usage. Lambrecht & Skiera (2006) applied nested logit to situations where consumers determine whether to keep or change their current tariff then, conditionally, the former could switch to another tariff of the same provider or churn; similarly, consumers in Wolk & Skiera (2010) choose an internet usage portfolio, and then, conditionally, a tariff. While the nested logit specifically alleviates IIA – and its potential for elasticity artifacts – overall, choice within a nest retains this property. However, Gu & Yang (2010), in studying nonlinear pricing for quantity discounts, adopt a flexible covariance specification specifically to alleviate IIA concerns, finding it superior to several nested logit structures. The proposed model will, in a sense, meld both these perspectives – plan selectivity and non-IIA conditional choice – and explicitly compare against restricted variances and nested logit.

The overwhelming majority of literature modeling plan choice stochastically relies on either a logit specification (entailing IIA), or a multinomial probit with either no error covariance or a tightly patterned one. For example, McManus (2007) modeled coffee size purchases at the University of Virginia, building a sophisticated structural model, but one in which size-specific errors were Gumbel distributed, allowing a tractable logit choice mechanism. Allenby et al. (2004) examined a budget-constrained, discrete-quantity framework and allowed for a full-covariance normal heterogeneity, yet choice was similarly via MNL. A similar presumption about lack of error covariance or correlated selectivity effects is adopted in much of the pricing and multi-part tariff literature, including those positing nonlinear utility

functions, e.g., Khan & Jain's (2005) analysis of consumer analgesics choice; Iyengar, Ansari & Gupta's (2007) learning-based model of wireless services; Goettler & Clay's (2011) study of grocery home delivery service; Ascaraza, Lambrecht & Skiera's (2012) and Grubb & Osborne's (2015) studies of cellular service; among numerous others.

The key observation is that the analyst wishes to understand substitution effects (i.e., elasticities) for all customers, particularly so those who did not self-select into purchase – after all, we wish to entice them to by altering the entire roster of plan prices – yet inferences made about them must rely on those who did. Such inferences must work off a statistical footprint that can be rather small, often (as in some of our experimental conditions) single percentage digits. Efficient methods for debiasing the resulting measurements must be devised, and such methods should carefully model the selection utility, the interrelated utilities of plan prices, and their mutual intercorrelations.

Before proceeding to the model proper, we first describe our field experiment in detail.

2.4 FIELD EXPERIMENT

2.4.1 Experimental Design

We implemented a field experiment in partnership with a U.S.-based online dating site[1] in February 2014. The site has operated since inception using a subscription model, offering multiple plans that differ in their contract duration (e.g., commitment length). Specifically, the site provides basic membership free-of-charge, allowing individual customers to create searchable public profiles as well as utilize a restricted set of site functionalities. Free users are fully aware of the benefits of upgrading to a paid membership, i.e., of capabilities of which they cannot avail. To access such paywalled 'premium' features (i.e., unlimited messaging, wider search area, etc.), a subscription plan may be purchased. Importantly, the functionalities themselves are not tiered, i.e., they are identical across paid plans: the only

[1] An NDA prevents disclosure of the site itself or information that might enable its identification. We point out elsewhere where specific information is deliberately redacted for this purpose.

consideration for users deciding whether to upgrade to a paid subscription is the length of contract to precommit.

The site offers three subscription plans, for one, three, and six months (hereafter referred to as "1MO", "3MO", and "6MO", respectively); these specific durations are standard for the firm and its competitors (e.g., Match.com, a market leader, also offers 1-, 3-, and 6- month plans). Upon joining, or at any time the user wishes to check, plan options are presented as a *price menu*, depicted in Fig. 2.1. The central conundrum for the user, as mentioned earlier, is a familiar one: longer contracts provide lower per-month prices, but at a greater overall cost paid upfront.

The site's executive team highlighted an underlying tension between higher monthly rates and locked-in revenue: while customers on the 1MO plan were the most lucrative per-single-month, those on the other plans afforded a longer guaranteed income stream. The firm explicitly prioritized the latter, and encouraged choice of the 6MO plan; as we shall see, this is the least popular choice both on their extant site and in in our experimental conditions. A key pragmatic concern was therefore understanding how altering the pricing menu (re)distributed customers (1) between free vs. paying; and (2) among the three subscription plans. To measure these core tradeoffs, we, in concert with the site's team, designed and implemented the randomized pricing experiment for *new registrants*, with the secondary purpose of relating demographics and other individual-level covariates to customer receptivity and profitability. The field data are especially rich in individual registrant record content, e.g., subscription service chosen, pricing treatment group ID, geographic location, mate-seeking preferences, among other situational and demographic variables that first-time users avail to the website upon registration.

2.4.2 Test Conditions

For an initial pooling period of one week, a randomly selected subset of new registrants to the site were funneled into the field experiment (N=18,286). To gauge uptake for subscription plans, the experiment ran for an additional 26 days; this timeframe was determined in conjunction with the site operators, whose historical analysis indicated that the vast majority of paid-subscriber first conversions occurred early in our experiment: 90% occurred by Day 12, 95% by Day 18, and 99% by Day 27.

Individuals in the experiment were randomly assigned to one of nineteen price menu treatment conditions (an additional twentieth condition was eliminated due to the firm's faulty implementation of its price levels). During the observation window (1-week intake + 26 days follow-up = 33 days total), subjects were exposed to pricing as determined by their treatment group assignment, and did not see any other pricing offers or promotions. All price offerings are viewed in a 3-option pricing menu analogous to Fig. 2.1, with trivial visual differences based on altering the price levels themselves. Participation in the field study was not disclosed to subjects.

As discussed earlier, real world prices in the online dating industry change rarely and, when they do, all plans tend to be raised simultaneously, making measurement of plan price cross-elasticities practically impossible. Our experimental design therefore sought to orthogonalize (1) the overall price levels of the trio of plans ("elevation"), operationalized via the 3MO plan price as the fulcrum, and (2) range of per-month prices ("steepness"), such that greater steepness made the 6MO plan relatively more attractive and the 1MO plan less so. Specifically, the 20 experimental conditions were based on a 5 (elevation) x 4 (steepness) design, which together characterize a pricing menu. [For the purpose of our subsequent analysis, "Base elevation, Flatish steepness" condition (Fig. 2.3) is omitted due to the aforementioned firm implementation error.] To ease unit-wise comparability, all prices hereafter are denoted as the per-month 'unit' price, unless otherwise noted. Because the study involved a large number of new registrants, the site was reluctant to include large price deviations from their standard price offering ('Base' = $18.99, Fig. 2.2), due to the possibility of substantial negative impact on revenue; there is no guarantee, therefore, that "optimal price menus" stemming from the analysis will lie in the convex hull of the experimental pricing levels. The firm was especially focused on measuring the effects of raising price, which they anticipated needing to do in the future; thus, the Base (B) price level (again, for the 3MO plan) was reduced by one dollar in the Lower (L) condition, but raised in the other three conditions: by $1 (H1), $2 (H2), and $3 (H3).

Steepness (Fig. 2.2) is the change in the per-month unit price of the three options, with the 1MO and 6MO prices represented as *multipliers* relative to the 3MO "elevation" price, which is especially relevant in forthcoming elasticity calculations. Steepness ranges from 140-170% on the 1MO multiplier and 65-80% on the 6MO, from the 'Flatish' (F) to 'Shallow' (W) to 'Standard' (D) through 'Steep'

(P) conditions, respectively. Note that 1MO and 6MO deviations from the baseline are asymmetric, with the former's percentage deviation being larger. As mentioned earlier, this asymmetry in steepness was imposed (by the firm) to limit deviations from industry practice pricing levels, in order to mitigate substantial revenue loss.[2]

To reiterate, "elevation" and "steepness" uniquely characterize each price menu, and orthogonalization allows us to distinguish their impacts on new registrants' subscription conversions, in contrast to the historical price time-series. Specifically, the elevation manipulation helps measure the absolute pricing effect between menus, while the steepness manipulation helps measure relative price effects within menus, both in terms of whether users subscribe and, conditional on that, which plan they select. Of particular importance is whether raising all prices (elevation) has muted effects compared with raising each plan separately, which could potentially lead to observed regularity violations (Tversky 1972). As detailed previously, because the chosen plan is observed only when ostensibly less price-sensitive users do decide to upgrade, selectivity artifacts can be induced, as discussed next.

2.4.3 Nonrandom Selection in Subscription Upgrade

The firm wishes to, in effect, measure a counterfactual: which plan would nonsubscribers pick *if* an array of plans were offered such that at least one that might appeal to them? Within the boundaries of site-based constraints and unwillingness to bear potentially nontrivial financial losses, the firm could therefore experiment with pricing plans to help determine substitution effects *among users who self-select into subscription.* This alleviates an important source of confounding present in data from the site's daily operation; namely, that those who subscribe have found a plan that represents an acceptable trade-off between costs and benefits. What it does not correct for is the self-selection itself: how can the firm extrapolate to the much-larger pool of users who chose not to subscribe during the data window? A question that is particular pertinent in attaining robust counterfactual simulations in optimizing prices across the entire cohort of new registrants, where different menus can result in upgrade cohorts that differ in their inter-plan substitution patterns.

[2] The firm further imposed that monthly plan prices be 'rounded' out to standard retail patterns, to end with ".99". All our conditions thus reflect this constraint.

In other words, assignment to one of our 19 conditions is random, but assignment into *subscription* vs. *non-subscription* is not. As discussed in the literature review, many models have been posited to "correct" for this non-random selection (Wachtel & Otter 2013; Feinberg, Salisbury & Ying 2016 provide recent reviews). Here, because we have full information (as detailed later) on all site users, we are able to extend these methods to the special case typical of service-based menu pricing: binary selection (into free vs. subscription) and multinomial choice (among plans). In the following sections, we develop the measurement model, and an efficient way to estimate its latent utility correlation parameters using Hamiltonian Monte Carlo (HMC; Neal 2011).

2.5 VARIABLE DESCRIPTION

Based on the data generated from the field experiment, we now describe the dependent and independent variable inputs for the empirical application of our model. The data consist of cross-sectional observations on the upgrade decision, conditional contract choice outcome, demographics, geographic, and mate-seeking preferences for the participants of our experiment. Overall, data from the 19 orthogonalized menu-price conditions are utilized, randomly assigned across 18,286 participants, of whom 3,758 (20.5%) subscribed within the experiment's observation period (as per Fig. 2.3).

2.5.1 Dependent Variables

The dependent variables of our model consist of a binary first-time upgrade decision, $Y_{s,i}$, and a multinomial conditional contract choice outcome, $Y_{o,i}\,|Y_{s,i} = 1$, at the participant-level, each of whom can be observed to have at most one $Y_{o,i}$. Among upgraders, we observed 2,383 (13.0%), 846 (4.6%), and 529 (2.9%) to have chosen the 1-, 3-, and 6- month options, respectively.[3]

The "Base elevation, Standard steepness" (Fig. 2.3) condition is the menu long-used by the site, and we note that it does not maximize yield for any of the three

[3] The 33-day observation window contained a few follow-up renewal decisions, but these were limited to participants who initially upgraded to the 1MO contract very early on. Such renewals were very rare ($< 1\%$ of participants), and don't correspond to *initial* decisions, and so are excluded from the analysis.

price levels. This is unsurprising, since it also does not offer the lowest prices for any of the three plans. Overall conversion is highest (23.7%), as might be expected, for the "Lower (L), Flatish (F)" condition, while conversion to the lucrative 6MO contract (5.08%) is highest for the "Base (B), Steep (P)" condition. Which condition is "best", of course, depends on how the firm balances short-term total gain (i.e., maximizing 6MO contracts), or high per-month fees (maximizing 1MO conversions). Our proximate goal is not to settle this particular issue, which is dependent on firm objectives, but how to measure price effect trade-offs across contract types in the presence of self-selection into conversion, although we do later optimize price menus for two such specific objectives, one suggested explicitly by the firm, and another utilizing historical CLV-based data.

2.5.2 Model-Free Evidence: Effects and Manipulation Checks

Because the experiment was operationalized via orthogonalizing "elevation" and "steepness", as a prelude to a model including individual-level factors into correlated latent utilities, we can examine the pattern of cell-wise experimental results in the 19 conditions (Fig. 2.3), for evidence of menu-price effects. To do so, we use median-splits, i.e., combining the "Lower (L)" and "Base (B)" conditions vs. the "Higher (H2)" and "Highest (H3)" conditions for the elevation manipulation; and the "Shallow (W)" and "Standard (D)" conditions vs. the "Steep (P)" and "Flatish (F)" ones for the steepness manipulation.[4] We can then compare, using likelihood ratio tests, what sorts of differences are observed in both selection and conditional choice. These tests, with the number of participants on which they are based, are given by Table 2.6.

As might be expected, "elevating" the price menu has a significant ($p = .015$) negative effect on subscription overall, from 21.2% to 19.4%; intriguingly, it also has no significant effect on the choice proportions across plans, although there is a mild directional effect upward on the 1MO plan (60.1% to 63.0%) and downward on the 6MO plan (16.7%). This is consistent with the idea that 6MO purchasers are the most price-sensitive, and differentially select out or to other plans when price shifts upward. Also intriguingly, "steepness" has a significant negative effect on subscription ($p < .001$), from 21.5% to 19.5%. We see that, as one might expect, the 3MO conditional choice rate is unaltered by this change – since its price is held

[4] Due to the implementation error in the "Base (B), Flatish (F)" condition, unbalanced cells must be omitted in each of these comparisons, which would upset the orthogonalization.

constant in absolute terms – but the 1MO rate is strongly affected downward (68.1% to 59.0%, $p < .001$), and the 6MO similarly strongly upward (10.3% to 17.4%, $p < .001$), suggesting that making the 6MO plan more attractive relative to the other two (and the 1MO less so) causes predictable shifts in the likelihood of choosing those plans. What one cannot tell from this pattern of results is what is driving them, that is, the effects of individual-level covariates and, more centrally, the plan-wise price substitution pattern, to which we return later in comparing cross-elasticities across models.

2.5.3 Menu Prices

To gauge the absolute and comparative effects of menu prices, we include linear, quadratic, and "ratioed" pricing terms into the utility specification. The linear and quadratic terms are consistent with previous work in nonlinear pricing (Iyengar & Gupta 2009), and serve to operationalize our experimental manipulation in the between-condition differences in elevation. The price ratio terms measure the differential impact of steepness, and consist of the unit price ratios between the 3- vs. 1- month contracts (3MO/1MO), 6- vs. 1- month (6MO/1MO), and 6- vs. 3- month (6MO/3MO). Note that across all price ratio terms, the shorter-termed contracts, which have the comparatively higher unit prices, enter into the denominator. Thereby all three terms are bounded between 0 and 1, such that conditions with lower duration discounts – i.e., "flatish" steepnesses – result in ratios approaching 1, and those with higher discounts/steepness give ratios approaching 0. These ratios are designed to capture relative within-menu tradeoffs due to pricing, decoupled from the absolute level effects captured by the linear and quadratic pricing terms. We introduce these sets of price terms into both the binary model for upgrading and the polytomous model for plan choice.

2.5.4 Demographic and Situational Variables

Table 2.1 lists summary statistics for demographic and situational variables, prior to mean centering and rescaling as inputs to the model. For categorical variables, in lieu of means, we report the number of categories. Furthermore, as all situational variables represent categorical optional self-reported fields on a user's public profile, each contains a "not reported" category, which we utilize as the baseline contrast for the variables' effects.

2.6 MODEL DEVELOPMENT AND ESTIMATION

To measure the effect of menu pricing on the (latently) correlated decisions of whether to subscribe and which plan to choose, we develop a *binary selection multinomial probit choice* model. The model is estimated using Hamiltonian Monte Carlo (HMC), a Markov chain Monte Carlo (MCMC) algorithm conducive to efficient sampling over highly correlated posteriors to recover of nonlinear parameters and full covariance matrices. We provide implementation details of our model estimation in both the general-purpose probabilistic programming platform, Stan (Carpenter et al. 2017), with source code provided in Appendix A.5.

Our framework provides an exact, full posterior generalization of the canonical (e.g., Heckman 1979) selectivity framework to discrete multinomial outcomes, which has long been acknowledged as an issue in empirical discrete choice applications (Dubin & Rivers 1989; Bushway et al. 2007). Common across the selectivity literature is a *bivariate* stochastic censorship mechanism governed by a common (correlation) parameter, ρ, which models degree-of-selectivity between the selection and outcome errors, and whereby the outcome of interest is only observed if the latent utility of selection is greater than some threshold (typically normalized to zero for identification). In generalizing to multinomial outcomes, we posit that selection censoring occurs through a *multivariate* normal (MVN) mechanism with a fully identified, positive definite error covariance matrix, which we formally introduce in the model development below. This flexible, matrix-based parameterization of the degree-of-selectivity enables the representation of each multinomial outcome with separate utility functions and outcome-specific correlation parameters.

Despite its potential broad applicability in both marketing and beyond, there has been no fully general analog of the Heckman method for polytomous choice outcomes. Our goal in developing the present model is to bridge the gap between extant selectivity methods and the multinomial choice setting typically faced by marketers, and to introduce a novel MCMC sampling strategy that exploits the information geometry of the model's marginal Hamiltonian dynamics to efficiently recover its parameters.

2.6.1 Model Development

In the classic Heckman (1979) framework, the response of an individual or consumer (i) is governed by a pair of correlated utilities with respect to selection (Y_s^*) and outcome (Y_o^*), where the outcome is only observed if the individual self-selects to respond ($Y_s = 1$). Formally, this is:

$$Y_{s,i}^* = X_{s,i}\beta_s + \varepsilon_{s,i}, \text{ where } Y_{s,i} = \boldsymbol{I}\{Y_{s,i}^* \geq 0\} \quad [2.1]$$

$$Y_{o,i}^* = X_{o,i}\beta_o + \sigma\varepsilon_{o,i} \text{ if } Y_s = 1$$

$$(\varepsilon_s, \varepsilon_o) \sim BVN(0, 0, 1, 1, \rho),$$

where in the case of dichotomous outcomes,

$$Y_{o,i} = \boldsymbol{I}\{Y_{o,i}^* \geq 0\},$$

such that both selection and outcome are marginally binary probit models. In extending this framework to multinomial outcomes, we recognize that the empirical context of our experiment observed *as-is* is composed of a pair of correlated choices: a dichotomous one on subscription upgrade, and a (conditional) polytomous one for contract. When selectivity is present, the corresponding random utility structure can then be specified as consisting of: (1) the $J - 1$ number of identified utilities for each of these two *choice categories*, and (2) a joint error covariance matrix across all utility functions to account for both within- and between- choice category correlations (Golob & Regan 2002; Zhang, Boscardin & Belin 2008).

The novel contribution of our framework is to address scenarios where polytomous outcomes that may be only partially observable, but where the inferential goal is to uncover substitution patterns from the full (uncensored) distribution. In particular, we wish to recover price elasticities on both the upgrade decision as well as contract choice across the full user base to robustly impute the profit-maximizing price menu(s) from our experiment. To do so, we consider multinomial outcomes *that are only observable conditional on the binary selection latent utility being positive* (i.e., selection censoring). In specifying our model, we note that the binary selection equation will be identical to that in Eq. 1. However, in the case of the censored multinomial outcomes, we observe $Y_o = j$ if both $Y_{o,j}^* =$

$\max\{Y_o^*\}$ and $Y_s = 1$, where Y_o^* represents the vector of latent utilities of J discrete outcomes:

$$Y_{s,i}^* = X_{s,i}\beta_s + \varepsilon_{s,i}, \text{ where } Y_{s,i} = \boldsymbol{I}\{Y_{s,i}^* \geq 0\}$$

$$Y_{o,i,j}^* = X_{o,i,j}\beta_o + \varepsilon_{o,i,j} \text{ if } Y_{s,i} = 1 \quad [2.2]$$

$$\text{where } Y_{o,i} = \max_j\{Y_{o,i,j}^*\},$$

$$(\varepsilon_s, \varepsilon_{o,1}, \dots, \varepsilon_{o,J-1}) \sim MVN(0, \boldsymbol{\Sigma}) \quad [2.3]$$

$$\text{where } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_s^2 & \rho_{s,o_1}\sigma_s\sigma_{o,1} & \dots & \rho_{s,o_{J-1}}\sigma_s\sigma_{o_{J-1}} \\ & \sigma_{o_1}^2 & \dots & \rho_{o_1,o_{J-1}}\sigma_{o_1}\sigma_{o_{J-1}} \\ & & \ddots & \vdots \\ & & & \sigma_{o_{J-1}}^2 \end{pmatrix}.$$

For identification, one alternative is held as the baseline to address the issues of additive redundancy. As a result, in Eq. 2.3, the error covariance term has only J-1 dimensions related to outcome errors. Similarly, to address the multiplicative redundancy of the latent utilities, the first elements in the error covariance matrix, with respect to both the selection and outcome submodels, are rescaled to 1, e.g., σ_s^2 and $\sigma_{o_1}^2$ (cf. Albert & Chib 1993; McCulloch & Rossi 1994). This typically occurs post-processing to produce canonically scaled coefficients from the estimation, which we discuss next.

2.6.2 Likelihood Function

For each individual participant in the field experiment, i, we observe a one-shot binary upgrade decision of whether to subscribe during the observation period (Eq. 2.1), and which of the J contracts to choose conditional on upgrade (Eq. 2.2). When upgrade does not occur ($Y_{s,i} = 0$), the contract outcome $Y_{o,i}\,|Y_{s,i} = 0$ is a conditionally censored outcome. However, we posit that a corresponding set of utilities $\{Y_{o,i}^*\}$ exist for each user regardless of the observability on their conditional contract outcome, representing their *latent* preferences toward the contracts, which exists for all users. Our estimation strategy is motivated and centered around the imputation of these latent values, which are characterized as *missing data*. To recover these, missing utilities, $Y_{o,i}^{*,mis}$, enter as individual-level parameters correlated

with $Y_{s,i}^{*}$ through the error covariance matrix, Σ (Eq. 2.3). As such, the parameter Σ captures the selectivity arising from multivariate stochastic censoring mechanism of subscription, and generalizes the "Heckman ρ" to polytomous choice settings. As derived in Appendix A.1, the full sample log-likelihood function is:

$$\mathcal{L}(\boldsymbol{\beta_s}, \boldsymbol{\beta_o}, \boldsymbol{\Sigma})$$

$$= \sum_{i=1}^{N} \sum_{j=0}^{J-1} d_{s,i} d_{o,i,j} \ln \left[\int_{E_{s,i} \cap E_{o,i,j}} \frac{1}{(2\pi)^{\frac{1+(J-1)}{2}} \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2} \varepsilon_i{}' \boldsymbol{\Sigma^{-1}} \varepsilon_i \right) d\varepsilon_i \right]$$

$$+ \left(1 - d_{s,i}\right) \ln \left[\int_{-E_{s,i}} \frac{1}{(2\pi)^{\frac{1+(J-1)}{2}} \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2} \tilde{\varepsilon}_\iota{}' \boldsymbol{\Sigma^{-1}} \tilde{\varepsilon}_\iota \right) d\tilde{\varepsilon}_\iota \right]$$

where: $d_{s,i} = \boldsymbol{I}\{Y_{s,i} = 1\}$; $d_{o,i,j} = \boldsymbol{I}\{Y_{o,i} = j\}$; $E_{s,i} = \varepsilon_{s,i} > -X_{s,i}\beta_s$; $E_{o,i,j} = \left[\left(\varepsilon_{o,i,j} > -X_{o,i,j}\boldsymbol{\beta_o}\right) \cap \left(\varepsilon_{o,i,j} - \varepsilon_{o,i,k} > \left(X_{o,i,k} - X_{o,i,j}\right)\boldsymbol{\beta_o}\right), \forall k \neq j\right]$; and $\varepsilon_i = \left(\varepsilon_{s,i}, \varepsilon_{o,i,1}, \ldots, \varepsilon_{o,i,J-1}\right)$, if $Y_{s,i} = 1$; $\tilde{\varepsilon}_\iota = \left(\varepsilon_{s,i}, \tilde{\varepsilon}_{o,i,1}, \ldots, \tilde{\varepsilon}_{o,i,J-1}\right)$, if $Y_{s,i} = 0$.

2.6.3 Estimation

As this likelihood is not amenable to closed-form computation, we employ full Bayesian inference to estimate our focal parameters, (β_s,β_o,Σ), using Hamiltonian Monte Carlo (HMC). The key challenge that our estimation strategy overcomes – which impedes a tractable approach for selectivity correction for multinomial outcomes – is the difficulty in efficiently sampling from the multivariate normal CDFs in the presence of high-dimensional, correlated missing data. Our strategy involves sampling the equivalent *data-augmented* (Tanner & Wong 1987) posterior over the set of latent utilities ($\mathbf{Y}_s^*$,$\mathbf{Y}_o^*$,$\mathbf{Y}_o^{*,mis}$), for both censored and uncensored outcomes, thus obviating the need to solve for the closed-forms of the integrals (details in Appendix A.5):

$$p(\boldsymbol{\beta_s}, \boldsymbol{\beta_o}, \boldsymbol{\tau}, \boldsymbol{\Omega}, \mathbf{Y}_S^*, \mathbf{Y}_O^*, \mathbf{Y}_o^{*,mis} | \mathbf{Y}_S, \mathbf{Y}_o, \mathbf{X}_S, \mathbf{X}_o) \propto$$

$$\propto \prod_{i=1}^{N} TMVN\left[\left(Y_{s,i}^*, \{Y_{o,i}^*\}\right) | \left(X_{s,i}\beta_s, X_{o,i}\beta_o\right), \Sigma\right] \cdot \prod_{d=1}^{D_S} N(\beta_s | 0,1)$$

$$\cdot \prod_{d=1}^{D_o} N(\beta_o | 0,1) \cdot \prod_{d=1}^{D_\tau} \text{Cauchy}^+(\tau | 0,1) \cdot \text{LKJ}(\Omega | 1) \cdot \left|\frac{\partial \tau}{\partial s_*}\right| \cdot \left|\frac{\partial \Omega}{\partial a_*}\right| \cdot \left|\frac{\partial Y_o^*}{\partial s_*}\right|$$

Moreover, Hamiltonian-guided data augmentation shows superior efficiency in terms of lower autocorrelation and higher effective sample size (ESS) as compared to more common Gibbs sampling approaches (Pakman & Paninski 2014). This is accomplished by incorporating curvature information with respect to the local manifold geometry of the posterior distribution when generating the sample Markov chain across iterations (Betancourt et al. 2017). As such, HMC is particularly attractive for sampling over augmented (e.g., latent utilities) and non-linear (e.g., error covariance) terms. We provide implementation details via the general-purpose probabilistic programming language Stan (source code in Appendix A.5).

For the estimation using our field experiment data, we obtain 2,000 posterior draws after a burn-in period of 500 iterations, across six chains. As in standard MCMC, which typically requires far greater number of draws (due to autocorrelation), convergence is assessed by the between-chain $\hat{R}$ statistic (Gelman & Rubin 1992), of which all parameters are found to be < 1.1.

Lastly, it's worth noting that our framework for correcting selectivity does not necessitate an exclusion restriction (i.e., instrumental variables). Analogous to a full-information maximum likelihood approach to the Heckman BVN selection model, our model identification arises from the nonlinearity of the MVN error structure (Li and Prabhala 2007, Wachtel and Otter 2013).

2.7 RESULTS

We present model parameters stemming from the field experiment data, first discussing key findings regarding the impact of menu-pricing and other covariates on first-time users' (1) decision to upgrade to the paid subscription service, and (2) choice between the 1-, 3-, and 6- month contract options. We then turn our attention to interpreting the presence of selectivity artifacts and their implications on the joint decision between upgrade and contract choice. In order to gauge the synthetic effects

of various modeling constructs on implied inter-plan substitution patterns, we compare our results – various model fit metrics and cross-plan elasticities – to those of three nested and three non-nested benchmarks.

Estimation results for the focal model appears in two parts: for the focal plan price covariates in both binary selection and conditional polytomous choice, along with latent covariance components (Table 2.2), and demographic and situational covariates. Recall that both the selection and conditional choice models include linear, quadratic, and price-ratio terms, as well as "all possible" (i.e., identifiable) latent utility correlations. Because all covariates are mean-centered, and prices are all in dollars, one can compare their values meaningfully. [Our discussion of "raw" coefficients will be brief, however; due to the highly nonlinear nature of the model, that all covariates enter into both selection and conditional choice, and that there are latent error correlations, discussion of marginal effects, in the form of elasticities, will be more directly illuminating, and appear in the following section.]

2.7.1 Binary Subscription Selection

With one exception (as below), posterior HDRs indicate all menu-price-related effects are very strongly significant. We find that the linear effects of price are positive for the shorter plans ($b = 0.250$, 1MO; $b = 0.265$, 3MO), but not for 6MO ($b = -0.123$), while all three plans have negative quadratic terms, consistent with a heightened distaste for higher prices: consumers differentially select out at especially high rates for any of the component prices increasing. Price ratios capture marginal substitution effects (on latent selection utilities) across plan prices; recall that these are set up as ratios of longer-to-shorter plans, so increases represent reducing the "duration discount". These are positive for 3MO/1MO ($b = 0.946$) and 6MO/1MO ($b = 0.590$), but ns for 6MO/3MO ($b = -0.207$). This pattern of results suggests that making the 1MO plan relatively inexpensive draws in more customers overall, consistent with the fact that this is the most popular plan, but which also may suggest that it serves as an "anchor" against which the other plans are evaluated. By contrast, adjusting the 6MO/3MO price ratio appears to have at most minor effects on selection.

The model contains many dozens of individual-level covariates, so we focus on several of especial interest. Among demographic variables, the premium service attracts customers who have fewer children ($b = -0.031$, $p < .01$), are younger ($b =$

-0.144, $p < .01$) and taller ($b = 0.030$, $p < .01$). Curiously, the number of potential partners in the user's geographic area was not a significant predictor of subscribing ($b = 0.005$, ns). We also found very strong effects for a number of self-reported categorical variables, such as body type, age of partner sought, education level, ethnicity, and religious affiliation, and much weaker effects for hobbies, and such appearance features as hair and eye color.

2.7.2 Contract Choice

Price menu parameter estimates for conditional plan choice appear in Table 2.2, using the 1MO option as the baseline for identification. In contrast to a three-option multinomial choice model applied to only users who subscribe, these estimates reflect how menu prices and other covariates influence plan choice if a first-time user were to upgrade, regardless of having been observed to do so. This in turn allows a "debiased" population-relevant counterfactual of menu-price changes on both upgrade and contract choice.

All "alternative-specific" coefficients are relative to the 1MO baseline used for identification, while "cross-sectional" (linear and quadratic) price effects are common across the plan alternatives. For these latter two, one sees a strongly positive negative price effect ($b = -1.431$), although the quadratic term is only marginally significant ($b = -0.345$), consistent with raising relative price decreasing choice probability, as would be expected.

In terms of plan-specific effects, both the 3MO ($b = -0.334$) and 6MO ($b = -0.366$) intercepts are negative, reflecting their being less popular than the 1MO plan overall. One would normally anticipate that all price ratio effects would have a complex pattern that depends on the extent to which the IIA property holds (conditional on subscribing). For example, for someone who has decided to subscribe, whether to choose the 1MO plan should not strongly hinge on whether the 6MO-to-3MO ratio is increased, over and above the price effects already accounted for (e.g., the linear and quadratic effects). Yet we consistently find this to be the case: The 3/1 price ratio has a negative effect on the 6MO plan and the 6/1 price ratio has a negative effect on the 3MO plan. Due to the complexity of these effects in the presence of selectivity, we defer a more "holistic" discussion for our examination of price elasticities.

Although the focus of our study is not demographic or situational variables per se, many of these do systematically seem to help predict contract. Just restricting to those demographics that seem to persuade users to select longer-term contracts (i.e., relative to 1MO), having children ($b_3 = 0.177$; $b_6 = 0.369$), wanting someone with the same smoking ($b_3 = 0.055$; $b_6 = 0.088$) or drinking ($b_3 = 0.214$; $b_6 = 0.196$) habits, and, curiously, not setting a minimum height ($b_3 = -0.140$; $b_6 = -0.391$) all have strong marginal effects on longer-term commitment, suggesting the potential to form targeting strategies based on the questionnaires users fill out upon initial sign-up.

More to the point, the pattern of menu price findings has much to tell the site, whose aim is not only to understand how the pricing menu affects upgrades of its user base, but to draw first-time users towards longer-term contract commitments. One might ask how the firm should alter the menu's elevation and steepness, and answering this question requires (as discussed earlier) specific assumptions regarding both discount rates and historical (re-)subscription patterns. We take up this question later in our price menu optimizations.

2.7.3 Latent Correlations in Selectivity and Plan Choices

The lower part of Table 2.2 presents estimates for the elements of the error covariance matrix that are not fixed by model identification. There are five of these, corresponding to the correlation between the selection model and the 3MO and 6MO plans; the correlation in conditional choice between the 3MO and 6MO plans, and the two diagonal elements for 3MO and 6MO error covariance scale. For our forthcoming elasticity model benchmark comparisons, we will set each of these, as sets, to their reference values (either 0 or 1) in order to better understand how they affect substitution patterns.

The primary econometric question we approach in this chapter is whether it's important to account for latent selectivity in contract choice. Accordingly, we focus on the selectivity terms, $\rho_{s,3}$ and $\rho_{s,6}$, which provide insight into the contract choice behavior of first-time upgraders, specifically, in the latent residual correlation between decision to upgrade and which plan is chosen. And we find this to be overwhelmingly so, with HDRs for both 3MO ($\rho_{s,3} = -0.361$) and 6MO ($\rho_{s,6} = -0.329$) very strongly below zero (classically, $p < 0.0001$). This pattern of results implies that, as the random component of the selection utility for users who upgrade

increases, their choice utilities for the 3MO and 6MO options (against the 1MO contract) decrease. These effects are over and above the "regressing out" of a very large number of individual-level covariates in both selection and conditional choice, whose effects can also be quantified.

One possible interpretation is that a first-time user who is curious but unsure of the usefulness of the premium functions may 'test the waters' by choosing the 1MO over the 3MO and 6MO plans, so as to avoid committing to the longer-term options. We stress again that deciding on specific plan prices is highly dependent on how the firm values near-term vs. longer-term revenue, a topic we return to at the end when we derive "optimal" price menus.

2.7.4 Model Comparison

It is useful to gauge the fit for the proposed model – which includes latent correlations between binary upgrading and multinomial contract choice – relative to changes in the model specification. Specifically, in ensuing discussions we refer to the proposed model as "Full" and consider three nested and three non-nested ones (common in the marketing and nonlinear pricing literatures).

Shown in Table 2.8, the "Full" model (M1) can be hobbled in terms of the main modeling innovation presented here: constraining the selectivity terms, $\{\rho_{s,3}, \rho_{s,6}\}$ to zero (M2); further by zeroing out inter-plan choice correlation ($\rho_{3,6} = 0$, M3), and even further by restricting choice disturbances to be equal ($\sigma_3 = \sigma_6$, M4). In terms of non-nested comparison models, we compare to three workhorses of discrete choice research: the multinomial probit (M5: MNP); McFadden's conditional multinomial logit (M6: MNL); and the nested logit (M7), where the explicit nesting structure involves initial contract choice. M5 can be viewed as a completely error unrestricted covariance structure, but one where "competition" between plans *and between subscribing at all* ("no choice") is governed by the same utility structure; M7 can be viewed as involving an initial selection step in the form of the nesting structure, but no error covariances among the plans themselves. And finally, M6 (MNL) is akin to a restricted hybrid in that there is no nesting (like M5) and no error covariance (like M7).

Table 2.3 compares M1-M7 in terms of a variety of overall fit metrics: in- and out-of-sample hit rate, log-probability (LP), and DIC (Spiegelhalter et al. 2002); each except DIC is also presented as a % change, with LP relative to correct

individual-level prediction across the $n = 18{,}286$ participants (i.e., $\exp(\Delta \text{LP}/n)$). Out-of-sample hit rates – on which we primarily rely for model comparison – are computed via 10-fold cross-validation: the full sample is divided randomly into 10 groups, each model is fit on 90% of the data, and the posteriors used to make predictions for the held-out 10%, by stochastically integrating over the latent utility covariances in models that include them. All hit rates for M1-M7 are computed across the entire posterior, and so "penalize" lack of parsimony. [We note that log-probability is calculated up to a data-dependent scaling constant, so values should be assessed only via *differences* across models; and also that DIC is known to be a potentially unreliable metric for selectivity models (Mason et al. 2012), so we do not use it directly for model comparison.]

A general idea of how the Full model performs can be gleaned from the "base" rates for each plan, comparing them to both in-and-out-of sample individual, posterior-averaged, predictions (Table 2.9). That is, a "base rate guess" at whether a randomly chosen participant subscribes to the $\{1MO, 3MO, 6MO\}$ plans would be correct $\{13.0\%, 4.6\%, 2.9\%\}$ of the time. The proposed model improves dramatically on these rates, especially for the less-common, longer-term plans.

Relative model performance (Table 2.3) can be assessed via hit rates in-sample, out-of-sample, and LP, and % changes based on each (base rates out-of-sample are also computed using tenfold cross-validation, by simply resampling observed outcomes). As might be expected, the Full model performs best across-the-board. A rough idea of which modeling constructs aid in prediction can be gleaned from comparing out-of-sample hit rates (OSHR). Using the "base rates", the Full model is on average nearly 18% better, but this is an easy benchmark to surpass. Of the non-nested models, the OSHR is best for M5 (MNP), which offers a similarly flexible covariance structure to the Full model but does not explicitly model selectivity separately; its OSHR is only 1.2% worse. By contrast, simply de-linking selection (M2) by zeroing out $\{\rho_{s,3}, \rho_{s,6}\}$ reduces OSHR by 5.2% (further restricting $\rho_{3,6} = 0$ and $\sigma_3 = \sigma_6$ appears to hurt in-sample, but not out-of-sample, performance). The performance of the other non-nested models is instructive: OSHR for the nested logit (M7) is 8.2% worse, and MNL (M6) 9.8% worse. Recall that both of these are common in the nonlinear pricing literature, with nested logit allowing some form of explicit account for the "branching" into subscription. However, its imposed pattern of latent utility covariance is apparently at odds with empirical data patterns. MNL

is a further restriction on the nested logit, and imposes the IIA property across all four options (1MO, 3MO, 6MO, no choice), and this further degrades predictive performance.

While it is difficult to summarize comparisons of metrics that account for different aspects of observed outcomes, the results of Table 2.3 do suggest that accounting for latent correlations substantially improves predictive performance, and that this is especially so for selectivity correlations. But predictive performance can be degraded substantially without necessarily skewing key estimates of pricing effects, which we take up next by examining cross-elasticity matrices for models M1-M7.

2.8 MENU PRICE ELASTICITIES

By their nature, model "parameters" are constants, and thereby assess a fixed marginal effect in some latent utility. However, they are notoriously difficult to contextualize in a complex model with literally hundreds of covariates and correlated "error" structures. As such, it is more useful to assess price substitution effects in a scale-free manner comparable across models, in the form of price elasticities. Such elasticities are, unlike the model's parameters, not constant; that is, they depend on "where" they are calculated. Because our focus is on moving the price menu from current practice toward ostensive optima (under the model), here we compute price menu elasticities at the current prices offered by the firm, \$29.99, \$18.99, and \$13.99 per-month rates for the 1MO, 3MO, and 6MO plans. We do this for the proposed model and the six others compared earlier. Recall that these benchmark models each fit less well, some far less so, both in- and out-of-sample; yet, because parameters can "adjust" to accommodate the nature of the data, it's entirely possible that their imposed patterns of error covariance – which can exacerbate IIA problems – may exert minimal effect on menu price elasticities. Recall as well that the nested logit explicitly accounts for the "initial" choice of whether to subscribe at all, while the MNP allows for a fully flexible error covariance, while suggesting that the self-selection into subscription is governed by the relative attractiveness of the "no choice" option compared with the three pricing plans. The goal is to seek out similarities and differences across them, to help determine which modeling constructs

and parametric restrictions may give rise to distortions of inter-plan substitution patterns.

The resulting elasticity matrices appear in Table 2.4, where we compute the effects of altering the three plan component prices on unconditional choice probabilities for each of the plans; we also compute the effects of the two main experimental manipulations, "elevation" (altering all plan prices in tandem), and "steepness" (moving the 1MO and 6MO prices in opposite directions). Results can be compared in terms of how "turning off" various parts of the model affects the cross-elasticity matrix; in the case of the nested models, these can be viewed as distortions caused by failing to account for various latent correlations and error scales between selection and plans, and among the plans themselves.

The "Full" model (M1) displays the classic elasticity pattern among plan prices, with negative values on the diagonal and positive on the off-diagonal. But the diagonal elements are quite different: apparently the 3MO plan is far more price-elastic ($e_{11} = -0.89, e_{33} = -2.68, e_{66} = -1.68$). This makes intuitive sense: customers with strong preference for a short- or long-term contract are less likely to substitute than those in the middle. It also makes sense that the 6MO plan is next down, since this is the least costly one (per month), suggesting those with highest price sensitivity might choose it to begin with. The cross-elasticities are similarly suggestive: raising the 3MO plan has almost no effect on choice of the 1MO plan ($e_{31} = 0.01$), and a small effect on the 6MO ($e_{36} = 0.22$), but quite a strong effect on the overall subscription rate ($e_{3s} = -0.57$). Overall, this suggests that raising the 3MO price causes come customers to switch to the 6MO plan, but many others to not subscribe at all. Contrast this with the 1MO plan ($e_{13} = -0.96$, $e_{16} = -0.76$, $e_{6s} = -0.22$), where some customers do apparently fail to subscribe, but there is clear substitution into the longer-term plans. The 6MO results are perhaps most interesting of all ($e_{61} = 0.42, e_{63} = 1.16, e_{6s} = 0.27$), where raising its price has a positive effect on overall subscription rates: since this is the least popular plan, and one chosen largely because it's inexpensive on a monthly basis, raising its price leads to a far higher take-up of the two shorter-term plans, which comparatively look like "better deals".

We also note that the "elevation elasticity" – where all plan prices are raised by the same small percentage – of -0.52 is actually a bit closer to zero than the value for just the 3MO plan on its own ($e_{3s} = -0.57$). While the difference is too small to

term is a full-blown regularity violation (Tversky 1972), it is intriguing that apparently raising price on the 3MO subscription alone has a similar effect to raising all three prices in tandem, consistent with the idea that consumers may be more forgiving about across-the-board plan price increases than individual ones. Lastly, the "steepness elasticity" measures small (local) changes to the 1MO and 6MO plans *in the opposite direction*, effectively increasing the range of prices. And, again, this effect is negative (-0.48) overall, but greatly hurting the 1MO plan, (-1.30), and strongly boosting the 6MO (2.44), while negligible effect on the 3MO plan (-0.19), as one might expect, given that its price is unchanged. The negative effect on subscription rate, however, suggests that the range of current prices might be too high in terms of attracting customers overall, but this would have to be coupled with a revenue analysis, to which we return later.

A key question concerns the effects of "turning off" the key modeling construct in this chapter: the two selection correlations, which were both very strongly significant. These results are for model M2 (in Table 2.4), where some of the elasticity values change dramatically. Most notable are the substitution effects for the 3MO plan, which are now all negative. Recall that the 3MO plan had by far the strongest own-price elasticity of the three for the Full model, and this is so for M2 as well, but those effects are in a sense "carried over" to the other plans, and the selection model – whose errors are no longer correlated with the choice model – struggles to account for the substitution patterns. As such, whereas $e_{3s} = -0.57$ for the Full model (M1), it is grossly inflated for M2, to $e_{3s} = -1.79$. Perhaps as a compensatory result, the overall subscription effects for the other two plans become much more positive: $e_{1s} = 0.14$, $e_{6s} = 0.61$. [Yet the "steepness" elasticities remain nearly unchanged.] Overall, given the decreased fit of M2 overall and the significance of the latent selection correlations, it would seem that excluding these – as is fairly standard in the literature – can produce nontrivially altered elasticity estimates. We note in passing that model M3, which turns off all latent correlations, evidences all the same distortions.

It is in a sense unsurprising that restricted versions of the focal model will fit less well, although the differences in elasticity patterns are nontrivial. One might ask the same of various non-nested models common in the literature, particularly so those in nonlinear pricing. M5-M7 are generally "well-behaved" in that they produce sensible cross-elasticities; but they may be too much so, attenuating some effects

and reversing others to have standard signs. Perhaps most notable is the 6MO elasticity on overall subscription (e_{6s}), which was positive in the Full model and all its restrictions, but which – perhaps owing to regularity – is negative for M5-M7, although it is quite close to zero for M7 (nested logit), suggesting that its ability to accounts for a type selectivity may help in this regard.

In terms of own-price elasticities, M5 (MNP) and M7 (nested logit) recognize that 3MO is the most price sensitive, while M6 (MNL) does not. This anomaly of MNL may owe to its "over-regularization" of price effects, wherein it imposes the greatest price elasticity on the least costly (per month) plan. Tellingly, it also suggests that each of the price effects on overall subscription has roughly the same value, around $e_{.s} \approx -0.3$ for all three plans, in stark contrast to all other models. Regardless, M5-M7 each suggests that the plan with the largest price elasticity on overall subscription is the 1MO plan, the one with the largest base of subscribers; further, M5 and M7 are strictly monotonic in this regard, with the 1MO plan the most elastic, the 6MO plan the least, and all negative. By contrast, the Full model and all nested variants identify the 3MO plan as most elastic, and a (mildly) positive effect of the 6MO plan.

2.8.1 Model-Free Evidence: Cross-Elasticities

One might question whether these results are mirrored in the raw condition outcomes – as listed in Fig. 2.3 – despite their lack of individual-level covariates, which ideally random assignment would help overcome (in effect putting these mean-shifter effects into the "error term"). Such an analysis is called to question by the small number of conditions, but it is possible to gain a crude sense of the patterns at play by simply regressing, for each of the three plans separately, the log-proportions of its subscription rate in the 19 conditions against the log-prices of each of the three plans. Doing so produces the following pattern of price plan cross-elasticities (Table 2.7).

Intriguingly, this (mostly) mirrors the results of the Full model: the 3MO plan is the most price elastic, followed by the 6MO plan; the cross-elasticities are very high for the 1MO plan; and the 6MO plan has a large positive substitution effect on the 3MO plan. Although one would not base pricing policies on such a simplistic, aggregate analysis, its confluence is nonetheless suggestive. We next turn our attention to fashioning actual price menu plans from the Full model.

2.9 PRICE MENU OPTIMIZATION

While the previous findings speak to the enhanced fit and flexibility of the proposed model over both nested and non-nested alternatives, its usefulness ultimately depends on its being amenable to optimization by firms utilizing price menus. Here, we carry out a full-scale price menu optimization relative to two revenue objective functions. First, a "myopic" one was suggested by the sponsoring firm to directly maximize total revenue at the time of plan choice. This effectively equates to assuming an infinite discount rate and thereby favors menus that induce choice on the 6MO contract, which provides the highest upfront earnings. We also consider a "long-term" horizon whereby we optimize over the expected annual revenue by accounting for the average resubscription rates for each plan choice, as calculated empirically from historical CRM data provided by the firm. As we shall see, this objective serves to diminish the importance of the 6MO in favor of the shorter-duration contracts. That is, while the 6MO contract may provide more initial revenue, the 1MO and 3MO plans have historically led to higher lifetime value. As noted by Wu, Zhang & Padmanabhan (2018), dating sites are unusual in that high customer satisfaction (finding a match) can manifest in higher churn; and if this is indeed the case with our partner site, then menu prices ought to be set instead to induce plan choice that maximize the expected per-period revenue over the duration customers remain with the site. Ultimately, our goal in undertaking this exercise is not to settle the specific question of whether the firm should prefer longer- vs. shorter- term revenue, but rather to demonstrate how our framework can be applied to support decision-making relative to a variety of firm-relevant objectives.

To optimize the entire price menu for a specific objective function, one can simply perform a suitably granular grid search, using the model's coefficients applied to the individual-level data for all 18,286 participants in the experiment. To stochastically integrate over the latent utility covariances, we draw 5000 IID random normal variates, process them through the Cholesky decomposition of the estimated covariance matrix, and thereby attain approximately 100M (i.e., 18,286 × 5000) simulated choices, so that accuracy in the expected revenues is on the order of five significant digits.

Fig. 2.4 presents the "marginal" curves for both long-term and short-term (myopic) expected revenues. Specifically, the simulation was carried out on a 50-cent grid over the \$0 - \$80 range for each menu price, with a finer 5-cent grid near

the overall optimized price menu for both objective functions. Note that, for each, there is a smooth, unimodal envelope for the best price for each plan holding the prices for the other two plans.

Recall that the firm is currently charging $29.99, $18.99, and $13.99 as per-month rates for the 1MO, 3MO, and 6MO plans. Under the Full model, the short- and long-term expected revenues at this set of price levels are $158,783 and $265,644, respectively (Table 2.5). If the firm adopts the "myopic" (short-term) optimal menu levels – thereby maximizing revenues "today" – the optimal price levels are $34.80, $19.15, and $17.75, resulting in expected revenues of $172,546, or an 8.7% increase. If the firm adopts a longer-term, CLV-based objective, the optimal menu prices are $35.50, $18.85, and $18.35, resulting in expected revenue of $296,050, an 11.4% increase (note that the revenue objectives have different horizons, with the former representing immediate commitment, so their values should not be compared).

In both cases, the model suggests their current prices are too low overall, by about 14.5% on average, nearly across-the-board (the exception being the $18.85 3MO price for the CLV objective), consistent with the firm's belief that they should be raised. In particular, the 6MO price appears much lower than it should be, roughly 30% overall, perhaps reflecting the firm's desire to maximize immediate revenue (since the 6MO plan requires by far the greatest up-front total commitment). Similarly, the overall range (akin to our "steepness" manipulation) is roughly 10% higher than it should be: the firm is apparently too focused on incenting customers into the 6MO condition, and would be advised to "compress" the price menu overall.

Comparing optima across the two objectives is illuminating. While the "elevation" (or average) price for the myopic and CLV objectives is roughly the same – $23.90 and $24.23, respectively – there is a substantial difference in their "steepness" between the 3MO and 6MO values, with the myopic one suggesting the 3MO be priced 8% higher, while the CLV-based one only 3%. This makes intuitive sense: the goal of the myopic objective is to slide subscribers into longer-term contracts. Yet these values are both well below that for the current prices (36%, i.e., $18.99 vs. $13.99), suggesting that the firm is overly incenting the 6MO contract, which is nevertheless by far the least popular choice. An analogous effect appears when comparing the 3MO vs. 1MO plan, which for the current prices are $18.99 vs. $13.99, or a 58% increase. Yet the model and optimization suggest it should be

substantially higher: 82% and 88% for the myopic and CLV objectives, respectively. In other words, the real message of the optimization appears to be that the firm is "leaving money on the table", and should simply raise price on both the 1MO and 6MO plans. In practice, the optimal prices would all likely need to hew to the "ends in .99" convention, but the directional results are suitably plain. In short, the model predicts that the firm is foregoing roughly 10% of its revenue using its current plan, regardless of adopting a short- or long- term perspective.

2.10 CONCLUSION

Duration-based menu pricing is ubiquitous in service plan choice, particularly so those priced "nonlinearly" on the web. Firms offering such services can attempt to rely on their own historical prices series to gauge the effects of component prices on both upgrading decisions and inter-plan substitution, but these tend to lack sufficient variation over time and plans, as well as perhaps being non-randomly triggered by, for example, failing to meet profitability targets. Here, we report on a menu pricing field experiment whose goal was to decouple the elevation and steepness effects that are ordinarily strongly confounded in menu pricing time series, as well as to fashion a comprehensive binary-multinomial selectivity framework for its analysis, one suitable for assessing price substitution effects in a wide range of nonlinear pricing models. Results indicate that model fit is substantially hampered by failing to systematically account for selectivity effects; that substitution effects can be biased by the error covariance assumptions baked into common discrete choice statistical frameworks for nonlinear pricing; and, substantively, that raising all prices concurrently can entail a smaller negative impact on subscription uptake than (some) individual prices alone. The model can also be used to optimize price menus over the entire range of their possible values. Doing so for our particular data set suggests that the firm is systematically underpricing its contracts – perhaps as a holdover from historical menu price levels – particularly so its "best deal", the longest-duration 6-month option.

While these results are at least partly in line with prior theorizing as well as model-free evidence of price substitution effects, they represent to our knowledge the first rigorous field measurements of both selectivity and menu pricing effects, as well as the use of highly-efficient HMC techniques to navigate the nonlinearities

intrinsic to the latent, cross-submodel error correlations, which here have a particular substantive interpretation. Still, the results can potentially go further along a number of dimensions.

First, the data and model are capable, as demonstrated, to engage in price menu optimization relative to various firm-relevant objective functions, this can depend on overlaying prior resubscription information. As such, firms must possess this information and, more important, be able to model how users resubscribe for each plan based on potential future price levels. That is, the firm must estimate how, for example, a user who finds the 6MO plan attractive in a particular price condition that has never been used before will resubscribe either at that price or an array of other ones down the road. Doing so would require a series of price experiments of the type carried out here, or simply waiting for such resubscription patterns to emerge, which can take quite some time to play out. Field experiments requiring multiple waves are notoriously difficult to conduct, due to privacy and attrition concerns, but a longitudinal orthogonalization of price menu values would be by far the most efficient and "clean" way to determine not only optimal price menus overall, but how to differentially market them to demographic groups.

Lastly, although our focal firm was not concerned when, within the data collection window, subscribers chose to do so, this is not always the case. That is, instead of a binary subscription model, the firm would wish to graft on a dedicated timing model to see whether, for example, intermediate, user-initiated usage predisposed certain uses to upgrade earlier during the experimental period. Such a framework could be set up using common hazard modeling techniques, if indeed the firm believes there was a strategic advantage to be had in gleaning this sort of information.

We see the core model, estimation methodology, and optimization framework presented here as readily extended to other subscription and menu pricing contexts, as well as polytomous choice scenarios impacted by selection at large, so long as suitable variation allows for parametric identification. The proliferation of intelligent, disintermediated online pricing algorithms should allow for a far greater variety of field tests in the future, of the sort that can allow for dynamic menu optimization as new data arrive.

2.11 TABLES AND FIGURES

Table 2.1 Summary statistics for demographic and situational variables

Variable	Description	Mean / # Cats	SD
AVAILMATCH	# of matches, per preferences	484.90	2200.57
SEX	Gender (Female = 1)	0.55	0.50
PHOTOS	Provided profile photo (Yes=1)	0.41	0.49
AGE	Age of user (minimum 18)	39.28	13.67
HEIGHT	Height, in inches	66.68	4.65
CHILDREN	# of children	0.35	0.89
LENGREETING	Word count, public greeting	202.81	331.39
LENSELF	Word count, self-description	60.17	18.19
LENOTHERTEXTS	Word count, optional free text	60.52	233.89
LANGUAGE	Language spoken (English = 0)	0.05	0.22
EDUCATION	Education level, categorical	10	
MARITAL	Marital status, categorical	4	
DRINK	Drinking habit, categorical	6	
SMOKES	Smoking habits, categorical	5	
ETHNICITY	Ethnicity, categorical	9	
RELIGION	Religion affiliation, categorical	25	
RELACTIVITY	Religious activity level, categorical	5	
BODYTYPE	Body type, categorical	9	
EYES	Eye color, categorical	10	
HAIR	Hair color, categorical	14	
PUNCTUAL	Punctuality, categorical	6	
TRENDY	Trendiness, categorical	6	
POLITICS	Political leaning, categorical	8	
OCCUPATION	Profession, categorical	9	
CORECOLOR	Site's "match type", categorical	5	

Table 2.2 Posterior summaries for price coefficients and error covariance

		M1		M2		M3		M4	
	Choice Alt.	**FULL**		**No Selection**		**No Correlations**		**Restricted σ**	
Binary Selection Model									
(Intercept)		-1.260	***	-0.908	***	-1.026	***	-0.610	***
Prices: Linear	1MO	0.250	***	0.105	***	-0.033	***	0.052	***
	3MO	0.265	***	-0.313	***	-0.101	***	-0.251	***
	6MO	-0.123	***	0.299	***	0.125	***	0.257	***
Prices: Quadratic	1MO	-0.487	***	0.016	***	0.011	**	0.018	***
	3MO	-0.059	***	0.006		0.013	**	0.004	
	6MO	-0.843	***	0.039	***	0.037	***	0.043	***
Price Ratios	3MO/1MO	0.946	***	1.417	***	-0.520	***	0.264	***
	6MO/1MO	0.590	***	-0.934	***	-0.929	***	-0.804	***
	6MO/3MO	-0.207	*	-1.271	***	0.590	***	-0.736	***
Polytomous Conditional Choice Model									
	Choice Alt.	**FULL**		**No Selection**		**No Correlations**		**Restricted σ**	
Cross-sectional									
Price: Linear		-1.431	***	0.184	***	-0.828	***	-0.118	
Price: Quadratic		-0.345	*	0.955	***	0.205	***	-0.088	
Alternative-Specific									
(Intercept)	3MO	-0.334	***	-0.129	***	-0.357	***	-0.129	***
	6MO	-0.366	***	-0.457	***	-0.276	***	-0.457	***
Price Ratio (3/1)	3MO	-0.149	***	-0.291	***	-0.029	**	-0.291	***
	6MO	-0.168	***	-0.279	***	-0.237	***	-0.279	***
Price Ratio (6/1)	3MO	-0.061	***	-0.205	***	-0.114	***	-0.205	***
	6MO	0.046	***	-0.144	***	-0.231	***	-0.144	***
Price Ratio (6/3)	3MO	0.104	***	-0.128	***	-0.171	***	-0.128	***
	6MO	-0.197	***	-0.131	***	-0.253	***	-0.131	***
Error Covariance Specification									
		M1		M2		M3		M4	
	Parameter	**FULL**		**No Selection**		**No Correlations**		**Restricted σ**	
Correlation Matrix									
	ρ_(s,1)	1	---	1	---	1	---	1	---
Selection: 3MO	ρ_(s,3)	-0.361	***	---	---	---	---	---	---
Selection: 6MO	ρ_(s,6)	-0.329	***	---	---	---	---	---	---
	ρ_(1,3)	1	---	1	---	1	---	1	---
3MO/6MO (within)	ρ_(3,6)	0.116	*	0.099	***	---	---	---	---
	ρ_(1,6)	1	---	1	---	1	---	1	---
Scale (Square Root of Diagonals)									
Selection	σ_1	1	---	1	---	1	---	1	---
Outcome: 3MO	σ_3	3.45	---	2.92	---	2.95	---	2.81	---
Outcome: 6MO	σ_6	2.94	---	2.68	---	3.06	---	2.81	---

Asterisks indicate that the highest posterior density (HPD) region does not contain zero at the 99% (***), 95% (**), or 90% (*) levels.

Table 2.3 *Fit Comparison Metrics for Full and Benchmark Models*

Model	Hit Rate (Individual-level)				Log Posterior Probability		DIC
	In-sample		Out-of-sample				
	Overall	% Change	Overall	% Change	Overall	% Change	
M0: Base Rates (whole sample)	65.1%	-16.9%	62.4%	-17.9%			
M1: Full Model	78.4%	--	76.0%	--	-9446.6	--	23977.4
M2: No Selection	77.2%	-1.6%	72.1%	-5.2%	-10314.9	-4.6%	25372.5
M3: No Selection or Correlations	74.6%	-4.8%	74.2%	-2.3%	-10222.1	-4.1%	34738.3
M4: No Selection, Correlations, Scale	75.7%	-3.4%	74.4%	-2.1%	-10292.7	-4.5%	44214.4
M5: MNP, Full Covariance	76.2%	-2.8%	75.1%	-1.2%	-12498.7	-17.1%	37016.3
M6: MNL (McFadden)	70.9%	-9.6%	68.5%	-9.8%	-11380.2	-10.5%	25972.8
M7: Nested Logit	71.1%	-9.4%	69.8%	-8.2%	-10910.7	-7.9%	26032.4

Table 2.4 *Menu Price Elasticities for Proposed and Benchmark Models*

FULL MODEL [M1]

	Sub%	1MO%	3MO%	6MO%
Proportions	19.7%	12.2%	4.4%	3.0%
Elasticities (Marginals)				
1MO	-0.22	**-0.89**	0.96	0.76
3MO	-0.57	0.01	**-2.68**	0.22
6MO	0.27	0.42	1.16	**-1.68**
Elevation	-0.52	-0.46	-0.56	-0.71
Steepness	-0.48	-1.30	-0.19	2.44

NO SELECTION [M2]

	Sub%	1MO%	3MO%	6MO%
Proportions	18.4%	11.2%	4.4%	2.8%
Elasticities (Marginals)				
1MO	0.14	**-0.59**	0.94	1.84
3MO	-1.79	-1.25	**-3.26**	-1.63
6MO	0.61	0.89	0.87	**-0.97**
Elevation	-1.05	-0.96	-1.46	-0.75
Steepness	-0.47	-1.48	0.07	2.80

NO SELECTION OR CORRELATIONS [M3]

	Sub%	1MO%	3MO%	6MO%
Proportions	18.3%	11.0%	4.7%	2.7%
Elasticities (Marginals)				
1MO	-0.09	**-0.92**	0.92	1.55
3MO	-1.29	-0.62	**-3.10**	-0.89
6MO	0.34	0.62	0.64	**-1.33**
Elevation	-1.04	-0.92	-1.55	-0.66
Steepness	-0.43	-1.53	0.29	2.87

NO SELECTION, CORRELATIONS, OR SCALE [M4]

	Sub%	1MO%	3MO%	6MO%
Proportions	18.0%	10.9%	4.2%	2.8%
Elasticities (Marginals)				
1MO	0.52	**0.28**	0.67	1.18
3MO	-2.49	-2.19	**-2.94**	-2.95
6MO	0.81	0.85	0.96	**0.38**
Elevation	-1.18	-1.06	-1.34	-1.42
Steepness	-0.29	-0.56	-0.29	0.78

Non-Nested Models [M5 - M7]

MNP, FULL COVARIANCE [M5]

	Sub%	1MO%	3MO%	6MO%
Proportions	15.2%	10.6%	3.2%	1.5%
Elasticities (Marginals)				
1MO	-0.64	**-1.05**	0.28	0.32
3MO	-0.35	0.15	**-2.32**	0.29
6MO	-0.14	0.05	0.09	**-2.06**
Elevation	-1.14	-0.85	-1.94	-1.47
Steepness	-0.50	-1.10	0.19	2.36

MNL (MCFADDEN) [M6]

	Sub%	1MO%	3MO%	6MO%
Proportions	20.6%	13.7%	3.9%	3.0%
Elasticities (Marginals)				
1MO	-0.37	**-0.99**	0.51	1.29
3MO	-0.22	0.03	**-1.67**	0.53
6MO	-0.31	0.03	0.03	**-2.30**
Elevation	-0.95	-0.95	-1.33	-0.45
Steepness	-0.08	-1.02	0.52	3.45

NESTED LOGIT [M7]

	Sub%	1MO%	3MO%	6MO%
Proportions	20.7%	13.1%	4.9%	2.7%
Elasticities (Marginals)				
1MO	-0.54	**-1.14**	0.65	0.21
3MO	-0.23	0.23	**-1.72**	0.22
6MO	-0.06	0.05	0.13	**-0.97**
Elevation	-0.79	-0.69	-1.01	-0.94
Steepness	-0.29	-0.95	0.38	1.72

Table 2.5 *Current vs. Optimal Price Menus and Revenue Projections*

	1 Month	3 Month	6 Month	Average	Range %	E[Myopic]	E[CLV]
Current	$29.99	$18.99	$13.99	$20.99	214%	**$158,783**	**$265,644**
Myopic	$34.80	$19.15	$17.75	$23.90	196%	**$172,546**	
% Change	16.0%	0.8%	26.9%	13.9%	-8.5%	**8.7%**	
CLV	$35.50	$18.85	$18.35	$24.23	193%		**$296,050**
% Change	18.4%	-0.7%	31.2%	15.5%	-9.8%		**11.4%**

Table 2.6 *Model-Free Evidence: Effects and Manipulation Checks*

	Subscription		Conditional Choice			
	# Custs	P(Sub)	# Custs	P(1MO)	P(3MO)	P(6MO)
	Lower vs. Higher Elevation					
Lower	5733	21.2%	1216	60.1%	23.2%	16.7%
Higher	5834	19.4%	1131	63.0%	22.8%	14.2%
p-value		0.015		0.158	0.827	0.100
	Lower vs. Higher Steepness					
Lower	7741	21.5%	1666	68.1%	21.6%	10.3%
Higher	7654	19.4%	1487	59.0%	23.6%	17.4%
p-value		0.001		0.000	0.181	0.000

Table 2.7 *Model-Free Evidence: Cross-Elasticities*

	1MO	3MO	6MO
1MO	-1.11	1.37	1.57
3MO	0.18	-3.53	-1.50
6MO	0.68	1.43	-1.93

Table 2.8 *Comparison Models*

		Restrictions	
	Model	**Correlations**	**Scale**
M1:	Full Model	---	---
	Nested		
M2:	No Selectivity Correlations	$\{\rho_{s,3}, \rho_{s,6}\} = 0$	---
M3:	No Correlations at all	$\{\rho_{s,3}, \rho_{s,6}, \rho_{3,6}\} = 0$	---
M4:	No Correlations or Scale	$\{\rho_{s,3}, \rho_{s,6}, \rho_{3,6}\} = 0$	$\sigma_3 = \sigma_6$
	Non-Nested		
M5:	MNP, Full Covariance	---	---
M6:	MNL (McFadden)	All = 0	All = 1
M7:	Nested Logit	No IIA	---

Table 2.9 *Hit Rate vs. Base Rate, In-sample and Out-of-Sample*

	1MO	**3MO**	**6MO**
In-Sample	34.3%	17.4%	14.8%
Out-of-Sample	31.6%	14.8%	7.1%
Base Rates	13.0%	4.6%	2.9%

Figure 2.1 *Price Menu Information As Displayed, Per Month and Total*

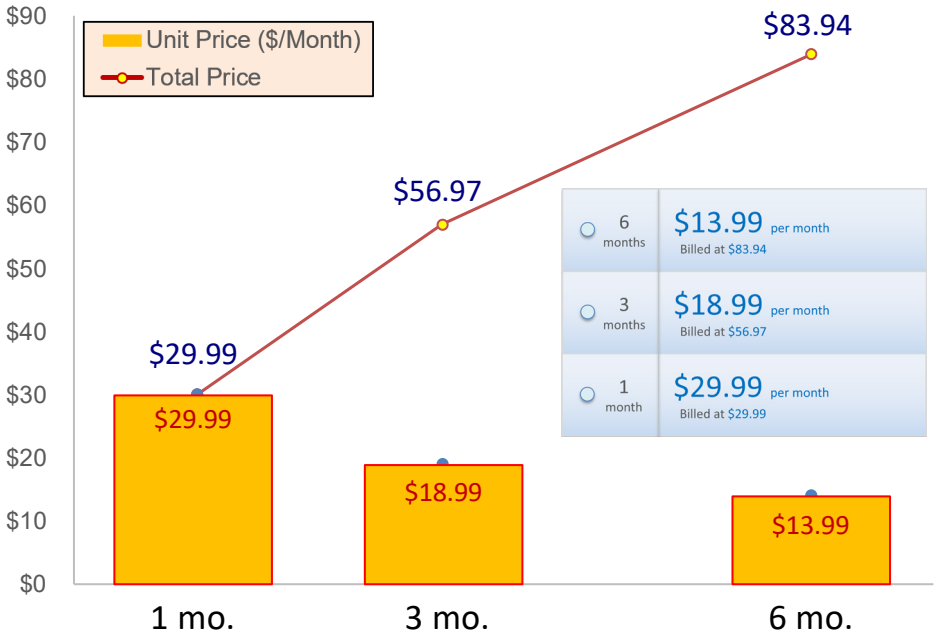


Figure 2.2 *Full Orthogonal Design*

Orthogonal 5 (price midpoints) x 4 (price gradient) design

Midpoints (3-Month Price)

Lower (L)	$	17.99
Base (B)	$	18.99
High (H1)	$	19.99
Higher (H2)	$	20.99
Highest (H3)	$	21.99

Steepness (multiplier on 3-month midpoint)

	1-Month	6-Month
Flatish (F)	140%	80%
Shallow (W)	150%	78%
Standard (D)	158%	74%
Steep (P)	170%	65%

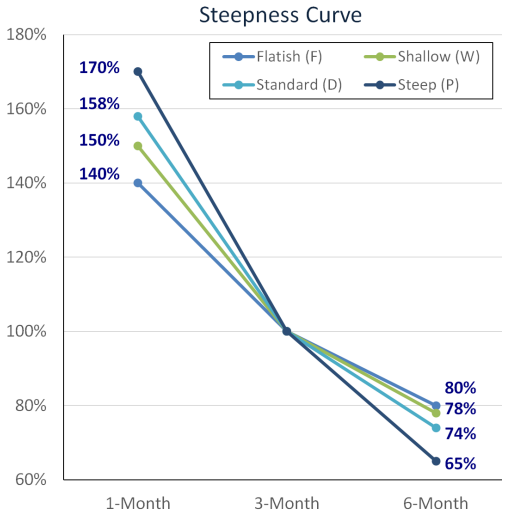


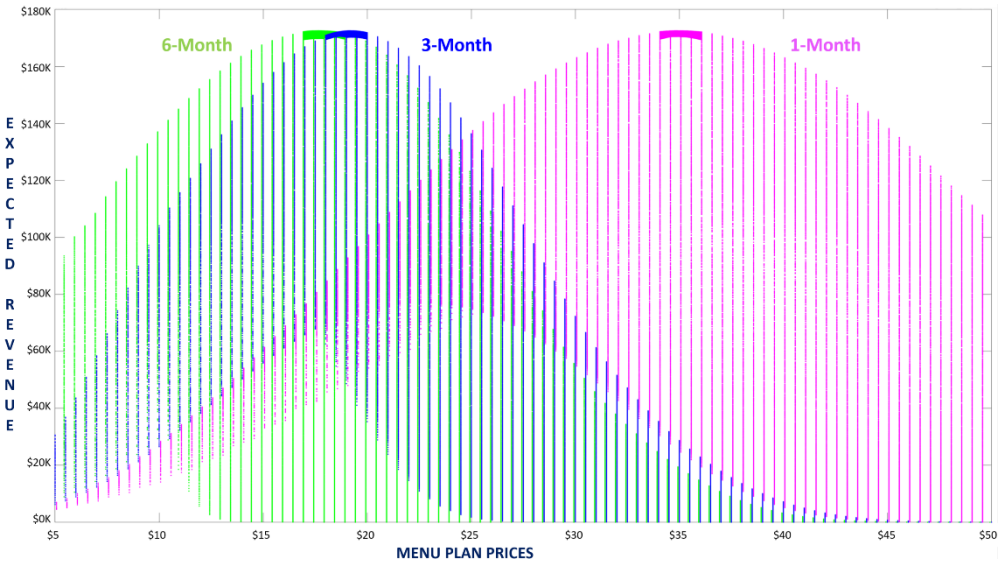
Figure 2.3 *Actual Conditions and Experimental Yields*

Elevation (Code)	Steepness (Code)	Unit Price ($/Month) 1 Mo.	3 Mo.	6 Mo.	Number of Users	Subscription Rates 1 Mo.	3 Mo.	6 Mo.	TOTAL
L	F	$24.99	$17.99	$ 13.99	934	15.8%	4.9%	2.9%	23.7%
L	W	$26.99	$17.99	$ 13.99	971	14.7%	4.7%	3.0%	22.5%
L	D	$28.99	$17.99	$ 12.99	918	12.9%	4.6%	3.5%	20.9%
L	P	$30.99	$17.99	$ 11.99	951	10.4%	6.2%	4.5%	21.1%
B	F	$26.99	$18.99	$ 14.99	ELIMINATED				
B	W	$28.99	$18.99	$ 14.99	991	12.4%	4.8%	2.4%	19.7%
B	D	$29.99	$18.99	$ 13.99	976	14.3%	5.2%	2.9%	22.4%
B	P	$31.99	$18.99	$ 11.99	926	11.7%	3.9%	5.1%	20.6%
H1	F	$27.99	$19.99	$ 15.99	1002	14.7%	4.2%	1.9%	20.8%
H1	W	$29.99	$19.99	$ 15.99	954	13.0%	5.5%	2.1%	20.5%
H1	D	$31.99	$19.99	$ 14.99	900	12.4%	5.1%	2.7%	20.2%
H1	P	$33.99	$19.99	$ 12.99	940	10.1%	3.7%	4.7%	18.5%
H2	F	$29.99	$20.99	$ 16.99	975	16.0%	4.4%	1.1%	21.5%
H2	W	$31.99	$20.99	$ 16.99	917	15.0%	4.9%	2.5%	22.5%
H2	D	$32.99	$20.99	$ 14.99	976	11.0%	4.4%	2.9%	18.2%
H2	P	$35.99	$20.99	$ 13.99	961	11.3%	3.5%	3.5%	18.4%
H3	F	$30.99	$21.99	$ 17.99	1016	15.6%	4.1%	2.0%	21.7%
H3	W	$32.99	$21.99	$ 16.99	972	12.4%	4.5%	2.3%	19.2%
H3	D	$34.99	$21.99	$ 15.99	1036	13.5%	4.5%	2.6%	20.7%
H3	P	$36.99	$21.99	$ 13.99	972	10.0%	4.6%	2.8%	17.4%

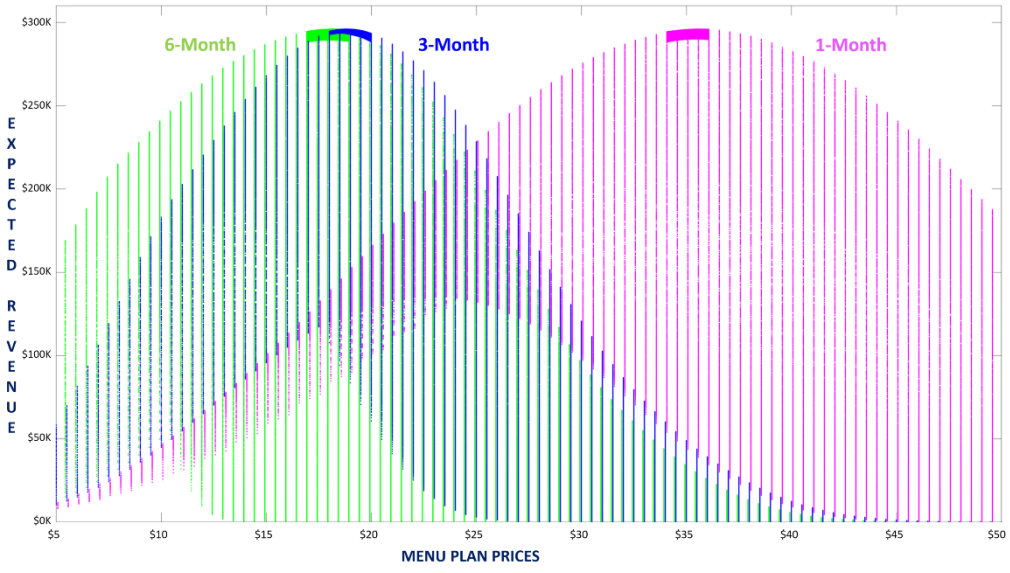
Note: Elevations are Lower (L), Base (B), High (H1), Higher (H2), Highest (H3). Gradients are Flatish (F), Shallow (W), Standard (D), Steep (P).

Figure 2.4 *Price Menu Optimization Contours and Expected Revenues*

A: Short-Term (“Myopic”) Objective



B: Long-Term (“CLV”) Objective



3 Broadening the Horizon: Augmenting One-Shot Field Experiments with Longitudinal Customer Data

3.1 ABSTRACT

Managers of online services are interested in both short- and long- term profitability, and how they are affected by customer-facing service attributes. This is especially true in subscription settings where different contracts, characterized by length, price, and other attributes, can affect long-run profits. A commonplace customer acquisition strategy utilized by these services involves offering price discounts targeted at new customers to induce initial conversion. As the initial contract choice of customers can have long-run impact on their subsequent renewal choices, setting new member pricing represents a tradeoff between running discounts steep enough to convert new customers while attracting customers into contract options that maximize customer lifetime value (CLV). However, problematic to conventional methods for measuring long-run effects of initial pricing, e.g., longitudinal A/B tests, is their time-consuming nature, as these require extended trial periods in order for long-run resubscription behaviors to play out.

To address this shortcoming, we introduce a Bayesian nonparametric data fusion framework that enables inference on the long-run effects of initial pricing using only a parsimonious 'one-shot' experiment on initial conversions, augmented with resubscription patterns found in longitudinal customer (CRM) databases. We apply our proposed framework to impute the long-run trajectory of resubscription choices for subjects from an A/B price test whose observations are limited solely to that of the initial conversions. We develop a class of Gaussian process (GP) prior data fusion models that utilize Bayesian regularization as the mechanism for the sharing of information across the datasets at the customer-choice -occasion level. The degree to which the longitudinal data regularize, or inform, the experimental subjects' likely renewal trajectories is given by the GP's Automatic Relevance Detection (ARD) kernel, which allows for differential degrees of regularization based on the distances between observations in the joint space of the customers' characteristics, price offers,

and time trends. Beyond the specific application to estimating the long-run CLV of the 'one-shot' experimental subjects, the data fusion framework can be generalized to any customer-base analytics employing a discrete-choice hazard function.

As GP computationally scales cubically with observations, $\boldsymbol{O}(n^3)$, conventional gradient- and MCMC- based estimation strategies are intractable given the large-scale nature of the fused data sources, both in terms of runtime and memory. To overcome these hurdles, we leverage the sparse 'inducing point' GP approach for Stochastic Variational Inference (SVI), in a first-of-its-kind application of this highly scalable Bayesian estimation strategy in marketing, with applicability to a broad class of choice, response, and latent class models.

3.2 INTRODUCTION

In crafting their pricing policies, firms face the ongoing conundrum of balancing long-run and short-term profitability. In practical terms, this plays out in setting levels to lure new customers with attractive rates while retaining existing ones at levels that ensure a healthy balance sheet. That is, they must differentially incentivize (groups of) customers at different stages of the customer lifecycle, e.g., from those newly acquired to eventual churners, at any given time.

For most services, particular those that utilize a subscription-based model, there typically exists a tradeoff on price-setting between the short- and long- term profit goals for a customer. In such cases, a common strategy is to offer upfront discounts or 'teaser rates' targeted at new customers that induce initial conversion at the cost of depressing short-run profits, while aiming to revert acquired customers to more profitable 'unpromoted' prices in the long-run. However, as suggested by Dube, Hitsch and Rossi (2010), pricing policy entails not merely a myopic focus on net-present-value of customers' initial contract choice, but forecasting effects on their subsequent resubscription behavior and, ultimately, on the core metric of customer lifetime value (CLV). The key challenge for practitioners and analysts alike is assessing these effects on CLV for newly acquired customers who, by definition, lack a detailed subscription history that would seem necessary to calibrate such a forecast.

In order to determine the price levels that optimize this tradeoff, firms require data that provide insight on how subscription and renewal decisions are affected by

prices and other customer-facing service attributes. Although customer relationship management (CRM) databases can contain a large 'breadth' of customers along with a 'depth' of recurrent subscription patterns, they are often plagued by low and non-random variation in key decision variables, such as discounts and promotional offers. On the other hand, firms, particularly those engaged in online services, can run longitudinal A/B tests, with randomized and orthogonalized pricing assignments, to assess the impact of introductory pricing on long-run resubscription choices and, thereby, CLV. However, such experiments can require extended time windows to allow sufficient renewal behavior to be recorded, compromising the ability to provide timely managerial decision-making.

To enable exactly this kind of CLV-informed decision-making, as mentioned above, firms can avail of two very different types of information: longitudinal data on existing customers that lacks systematic, controlled variation in inputs of interest, and 'one-shot' experiments on newly acquired customers that lacks downstream renewal behavior. To leverage the differential strengths of these commonplace data sources to jointly overcome their individual shortcomings, we introduce a Bayesian nonparametric data framework for "fusing" field experiments and CRM databases to undertake longitudinal customer-base analyses that would be difficult or misleading with either of these data sources alone. In an application to an online subscription service, we investigate the CLV effects of initial pricing plan using data from a parsimonious 'one-shot' price experiment whose observations are limited solely to that of the initial conversions, augmented with longitudinal subscription patterns from a large-scale CRM database.

One specific goal in fusing onto longitudinal CRM data is to uncover the likely trajectory of future subscription behaviors of the subjects found in the one-shot experiment given their initial pricing treatment. Core to our data fusion framework is the enabling of robust inference on a class of dynamic heterogeneous models for customer-level repeat-choices based on the multinomial logit (MNL) and probit (MNP) specifications. Our proposed models utilize the Bayesian nonparametric Gaussian process prior (GPP) as the mechanism for the sharing of information across the datasets at the customer-choice-occasion level. The degree to which the longitudinal data regularize, or inform, the experimental subjects' likely renewal trajectories is given by the design of the GPP's Automatic Relevance Detection (ARD) kernel, which allows for differential degrees of regularization based on the

distances between observations in the multidimensional joint space of the customers' characteristics, price offers, and time trends, which we refer to as the customer time-space. As such, beyond our specific application, the framework is generalizable to any customer-base analysis involving temporal discrete choice outcomes within a contractual setting, including the measurement of acquisition propensity, churn rate, and purchase outcomes.

In undertaking data fusion between the experiment and CRM database, we posit a commonly occurring data generating scenario faced by online subscription services running one-shot A/B tests on new users: upon the conclusion of the trial period, the experimental subjects who were previously exposed only to their treatment assignment now enter into the common pool of customers, e.g., those found in a CRM database. Standard practices dictate that to obtain any long-run estimates on an outcome of interest with respect to the initial manipulations, the firm should simply wait for the desired length of time before taking measurement. We aim to expediate this measurement by exploiting the very fact that upon the conclusion of the one-shot trial period, test participants are no longer subjected to controlled conditions but rather have now "become" customers in the CRM database. As such, any long-run inference on initial treatment beyond the one-shot test period are then privy to pricing and other marketing mix regimes akin to other CRM customers. Measuring long-run effect of initial treatments with intermediate non-controlled variations is prevalent to marketing field studies (Danaher 2002, Anderson and Simester 2004). Our research extends this literature by seeking the counterfactual of how the one-shot test subjects would behave in the long-run, given that they become a database customer upon the conclusion of the trial period? In addressing this question, intuitively, our proposed method is to augment the experimental data with longitudinal renewal patterns already observed in the CRM database using data fusion, which in turn allows inference on the desired long-run effect in an expedited fashion.

This study extends prior data fusion work by allowing for the first-time nonparametric dynamic evolution in customer-level preferences when augmenting experimental and survey data with observational data for the purpose of robust inference on heterogeneous response (Feit et al. 2010, Feit et al. 2013). Previous efforts in Bayesian data fusion have tended to rely on hierarchical specifications that necessitated parametric restrictions on the data fusion mechanism as well as

precluded the ability to capture individual-level time dynamics. Building on the nonparametric fusion approach of Qian and Xie (2011, 2014) to obviate the need for distributional assumptions on the fused variables (Raghunathan and Grizzle 1995, Rässler et al. 2002, Adigüzel and Wedel 2008), we employ the Bayesian nonparametric GPP to automatically infer suitable nonlinear functional representations for the evolution of choice preferences across customers and choice occasions in the combined data. While the ARD-kernel GPP is utilized in this study as a nonparametric data fusion approach, the proposed framework presents a unifying nonparametric approach for providing flexible, multidimensional shrinkage to a broad class of choice, response, and latent class models in marketing.

The use of GPP also allows for relaxing the conditional independence assumption (CIA) typically necessary when undertaking nonparametric and likelihood-based data fusion approaches, where a fully generative parametric model over all data is not specified. CIA arises from the perspective that data fusion is a missing data problem (Little and Rubin 2002) where the goal is to enable inference on the joint distribution of target variables that are not directly observed together, but whose marginal distributions may be imputed from separate data sources that have shared common variables (Gilula et al. 2006). As the target variables are imputed through 'matching' on shared variables, then matched variables are assumed to be conditionally independent (Rubin 1973). Relaxing this assumption has been a central focus among recent data fusion research. Although an argument can be made for CIA in data where an informative set of common variables are used (Qian and Xie 2014), it is problematic in our context where the goal of data fusion is to uncover time-varying preferences over and above those captured by shared observables and parameters.

Our approach to alleviating CIA is three-fold. First, taking the missing data perspective, the CRM database empirically contains the desired joint variation of observables and subscription choices over time, whereas only the 'one-shot' experiment is missing observations on its subjects' subsequent renewals. Hence, data fusion in this context can be seen as the process of augmenting the experimental data with the joint temporal variation found in the database. In a related study, Gilula and McCulloch (2013) demonstrate that this provides superior estimates than fusing data sources where the joint density of the target variables is entirely unobserved. Second, the GPP data fusion models are specified with a full-covariance

structure across all customer-choice-occasions in the combined data. Despite the obvious model-based advantages to a full-covariance term, making the CIA in prior literature is in part motivated by the need for computational tractability. We overcome this hurdle by employing the sparse 'inducing point' Gaussian process approach (Titsias 2009) for stochastic variational inference (Hoffman et al. 2013), in a first-in-kind application of this highly scalable Bayesian estimation strategy for marketing data sources. Lastly, in line with prior full-information Bayesian data fusion approaches, posterior inference on the augmented missing variables draws upon not only shared covariates (i.e., observed variables) but also shared parameters (i.e., latent variables).

This chapter is organized as follows. In the next section, we provide details on the Bayesian nonparametric data fusion modeling framework. In section 3.4, we discuss the empirical context to which we apply our framework, along with model-free evidence. In section 3.5, we describe the scalable inference strategy based on stochastic variational inference. In section 3.6, we present preliminary results of this ongoing study. Lastly, we conclude with managerial implications of our methods and future direction of this study.

3.3 MODEL DEVELOPMENT

We now provide details to the development of the Bayesian nonparametric data fusion model. In our application, we focus on augmenting the one-shot subscription price experiment with longitudinal database records to estimate a multi-period CLV model for subjects of the experiment. However, the data fusion framework we provide here can be applied at large to any temporal customer-base analysis employing a discrete time hazard function whose per-period choice and survival probabilities are estimated from a dynamic choice model, including the measurement of acquisition propensity, churn rate, and purchase selections. To this end, we provide derivations of our data fusion framework based on both the multinomial logit (MNL) and probit (MNP) specifications, given the prevalence of these choice models in the literature and among industry practices in marketing.

The CLV model that we seek to estimate is at the individual customer-level (denoted by $i = 1, \ldots, N$). Beginning with the customer's initial subscription choice at $t = 1$, we observe T sequential subscription-choice occasions where her choice

Y_{it} is among the set of J subscription options (e.g., 1-, 3-, 6- month, and no plan choice), for which each has a corresponding (promoted) price of p_{ijt}, and marginal cost of c_{ijt}:

$$E(CLV_i) = \sum_{t=1}^{T} \sum_{j=1}^{J} \frac{\Pr(Y_{it} = j) * \left(p_{ijt} - c_{ijt}\right)}{\delta(t)} \qquad [3.1]$$

As the nominal unit of time t is unitized across subscription-choice occasions (e.g., initial subscription and subsequent renewals), the discount factor is given by a function $\delta(t)$ to denote that the discount rate should be a transformed calculation with respect to the corresponding calendar time[5] that will have passed since the initial upgrade choice at occasion $t = 1$, or alternatively, with respect to the analysis's point-in-time. Although the total number of choice occasions T can be arbitrarily large, it should be *a priori* designated by the analyst in consideration of the available (longitudinal) data that could empirically identify long-run choices without being statistically underpowered.

3.3.1 Calculating Choice Probabilities: Experiment vs. Database

At the heart of our CLV model (Eq. 1), and the focus of our modeling framework, is the individual-level per-period choice probability, $\Pr(Y_{it} = j)$. The need for a data fusion approach in estimating Eq. 1 for the subjects from a one-shot experiment is motivated by the lack of observations beyond $t = 1$, hampering the estimation of $\Pr(Y_{it} = j)$ for subsequent renewal occasions. A naïve approach is to do away with time dynamics and assume that for all $t = 1, \dots, T$:

$$F^{-1}(Y_{it} = j \mid p_{ijt}, X_{it}, \Theta) = \alpha_j + p_{ijt}\beta + X_{it}\gamma_j + \varepsilon_{ijt} \qquad [3.2a]$$

where F^{-1} is the canonical link (or inverse cumulative density) function of the parametric choice model (e.g., MNL, MNP), X_{it} and γ_j are covariates related to individual i and their alternative-specific coefficients, and Θ denotes the set of all model parameters. Here these include: α_j, the choice-specific intercept; and β, the price effect across all subscription alternatives. Note that although the one-shot experiment is inherently cross-sectional without future observations on the

[5] The derivation of $\delta(t)$'s functional form is trivial as we assume a calendar-time exponentiated discount rate in our exposition, although practitioners may choose to use more sophisticated discounting strategies and rates, for which Eq. 3.1 can be readily 'plug-in'.

subjects, we denote subscription-choice occasion covariates with a time-subscript (X_{it}) to indicate the inclusion of static features (e.g., gender, self-description, physical features) along with extrapolatable time-varying features. The latter include membership duration, prior subscription choices, and other covariates that can be extrapolated or carried over from prior periods. Estimating this model (Eq. 3.2a) on the data generated from the orthogonalized and randomized experiment is expected to provide an accurate measure of subscription-choice tradeoffs β (and by extension, the other coefficients) during initial upgrades. However, if such a model were to be estimated as described here, it's unlikely to accurately impute CLV beyond $t = 1$ as the data do not inherently avail this information, and thus, any predictions on renewals is enabled by model assumptions alone (i.e., no time dynamics). We will utilize this 'naïve' experiment-only model as one of the benchmarks for our focal model.

Alternatively, the analyst may choose to leverage the expansive history on subscription choices typically found in a longitudinal CRM database. With this data, time dynamics and correlations can now be introduced both parametrically (e.g., α_{jt}, correlated $\{\varepsilon_{ij}\}^T$), and through additional observables X_{it} (e.g., past plan choices, membership duration, time-trend fixed-effects, etc.)[6]:

$$F^{-1}(Y_{it} = j \mid p_{ijt}, X_{it}, \Theta) = \boldsymbol{\alpha_{jt}} + p_{ijt}\tilde{\boldsymbol{\beta}} + X_{it}\gamma_j + \varepsilon_{ijt} \qquad [3.2b]$$

The key distinctions between Eqs. 3.2a and 3.2b are that (1) the latter's alternative-specific intercept is time-varying across choice occasions, and (2) the subscription-price effects identified from the database no longer arise from price variations of an orthogonalized RCT, and consequently, we denote this by $\tilde{\beta}$ to highlight this difference. Note that for (1), we do not additionally include an individual-specific intercept (e.g., α_{ij}), as it is rare for individual customers in our context to have enough repeat choices in every subscription option to meaningfully identify the effect (cf. Feit et al. 2010). For (2), the firm has historically offered promotions that only *contemporaneously* modulated prices (e.g., 20% off all subscription plans), followed by long periods of static 'standard' unpromoted pricing. This is in contrast to the fully orthogonalized variation provided from the

[6] These time-based covariates are in principle available within the experiment, although are likely to exhibit less variation than the database given a overall shorter observation time window.

experiment (i.e., Eq. 3.1). As a result, estimating $\Pr(Y_{it} = j)$ from Eq. 3.2b using the longitudinal database entails that the marginal probabilities $\partial \Pr_j / \partial \tilde{\beta}$ are no longer robust to the assumption of "holding all other prices constant". A critical and undesirable consequence of employing Eq. 3.2b that's fitted on the longitudinal database is that without the exogenous variations as described, any price optimization requiring *ceteris paribus* marginals cannot be presumed reliable to customers' subscription option tradeoffs and therefore their choice outcomes.

Along with the experiment-only model (Eq. 3.2a), the database-only model (Eq. 3.2b) will serve as another benchmark comparison to our data fusion model, which we now detail.

3.3.2 Data Fusion: Joining Experimental and Database Variations

The purpose of data fusion is to enable inference on the joint distribution of variables that are not directly observed together (Gilula et al. 2006), but whose marginal distributions may be imputed from separate data sources that share overlapping common variables (Kamakura and Wedel 1997, Qian and Xie 2014). It is then from the individual datasets' concurrent marginal covariances with respect to the *shared variables* that the joint distribution of the target variables is inferred. Bayesian data fusion extends upon this framework for inference on the joint distribution by also taking into account *shared parameters* between the datasets (Feit et al. 2010).

Data fusion methods can be broadly classified as taking a direct estimation (DE) or multiple imputation (MI) approach. In MI, fusion is conceptualized as a missing data problem (Rubin 1986, Feit and Bradlow 2016) whereby target variables are imputed through 'matching' on shared variables. Underlying much of earlier works in this area, the conditional independence assumption (CIA) of the target variables[7] is a result of the perspective that matched variables are conditionally independent (Rubin 1973). Relaxing this assumption has been a central focus among more recent data fusion research, particularly those on DE. Gilula and McCulloch (2013) proposed an empirical Bayes strategy to alleviate CIA for fusing categorical data where a set of evidence-based dependency rules serve as priors on the cells within the two-way table of the categorical targets. Feit

[7] $f(A, B|C) = f(A|C)f(B|C)$, where A and B are the target variables to be fused and C denotes the 'common' shared variables.

et al. (2010) proposed a DE likelihood-based fusion technique using hierarchical shrinkage to account for dependency or – taken from a Bayesian perspective – the sharing of information, across datasets. Although MI is computationally straightforward and can avoid restrictions on the form of the target variable's conditional distribution, more complex models such as those involving heterogeneous and hierarchical specifications are only feasible through a 'joint' likelihood approach across datasets (Feit and Bradlow 2016).

Our fusion framework takes a DE approach as our goal is to uncover dynamic preference heterogeneity for subjects in a one-shot experiment upon augmentation with longitudinal data. We also address the need to alleviate parametric restrictions on the fused joint distribution through the use of Bayesian nonparametrics to flexibly capture the joint distribution of the target variable-of-interest. In comparison to Qian and Xie (2011, 2014) where a nonparametric odds-ratio fusion is employed within an iterative procedure alternating between data imputation and parameter estimation, our proposed framework avoids the need for imputation, and instead directly integrates data fusion into the likelihood of the choice models. This streamlining enables the use of stochastic variational inference (Hoffman et al. 2013), an efficient and scalable modern inference technique, to overcome the 'small data' limitations of existing data fusion methods that would have otherwise hampered our ability to make use of the entirety of the longitudinal database (> 1 million observations) in our application.

Extending Eqs. 3.2a and 3.2b, the focal innovation of the proposed Bayesian nonparametric fusion framework is to enable inference on the dynamic individual-level alternative-specific intercept (α_{ijt}) that would otherwise not be possible on the one-shot experimental dataset alone:

$$F^{-1}(Y_{it} = j \mid p_{ijt}, X_{it}^{E}, \{X^{L}\}, \Theta) = \boldsymbol{\alpha}_{ijt} + p_{ijt}\beta + X_{it}\gamma_j + \mu\varepsilon_{ijt} \qquad [3.3a]$$

$$\boldsymbol{\alpha}_{ijt} \mid X_{it}^{E}, \{X^{L}\} \sim \mathcal{GP}(\cdot) \qquad [3.3b]$$

where μ is the between-dataset scaling factor[8] for discrete-choice fusion (Swait and Louviere 1993), X_{it}^{E} denotes the static and time-extrapolatable covariates of

[8] Our notation here for the discrete-choice data fusion scaling factor is given in terms of two datasets. More generally, for fusing D datasets, the scaling factors have the constraint $\sum_i \alpha_i \log(\mu_i) = 0$, with $\sum_i \alpha_i = 1$.

the *experimental* subject, and $\{X^D\}$ as the corpus of longitudinal customer data from the fused CRM *database* where information on the likely long-run trajectory of the individual preference for alternative j is derived. The identification of α_{ijt} (Eq. 3.3b) is made possible by through a data fusion *Gaussian process prior* (GPP) specification that enables the augmentation of the experimental subject i's data (X_{it}^E) with renewal information of similar customers from the longitudinal database $\{X^L\}$. We provide details on the specification of the GPP in the next section.

Although Eqs. 3.3a and 3.3b are given in relation to the end goal of estimating the long-run CLV of the experimental subjects, the proposed framework wholly fuses both the experimental and database observations. As such, in line with prior work in data fusion, estimates on (β, γ_j) leverage variations from the fully fused data to improve inference on customers from both sources. In particular, we denote the fused price-effect as β as it is identified on both the orthogonalized variations from the experiment as well as the contemporaneous price-change-only variations from the database (denoted earlier as $\tilde{\beta}$). Other effects such as seasonality and time-trends (elements of γ_j), which were limited in their generalizability due to the narrow observation window from the experiment (33 days), can now be augmented by variations arising from the longer time horizon of the database. Similarly, the Gaussian process over α_{ijt} is estimated on the fused data and therefore inferable for all customers, and as a comparison, we compute the marginal predictive accuracy of models leaving out such fusion (Eqs. 3.2a and 3.2b) using a hold-out sample.

Note that the fused model intercept (α_{ijt}) and error term (ε_{ijt}) are both unitized at the customer-alternative-choice-occasion level (ijt). Their distinction lies in that across all observations, the former represents the customer-choice-occasion specific baseline preference, and the latter captures the unobserved component of the utility function. This distinction is reflected in their prior specifications, whereby the former is given by the GPP and the latter is based on the standard corresponding random utility error specification of the chosen discrete choice model (e.g., Type I Extreme Value for MNL, zero-mean multivariate normal for MNP). Moreover, although only the intercept is specified to be dynamic among the focal parameters $(\alpha_{ijt}, \beta, \gamma_j)$, time-varying covariates are included in X_{it}. As such, our model specification can be interpreted as having (β, γ_j) capture the main-effect on subscription choice, and α_{ijt} capture the remaining time-varying

preferences of the individual customer. It is therefore imperative that α_{ijt} is simultaneously flexible with respect to complex time-evolution of preferences while being well-identified at the individual-level. To accomplish this, we fashion a novel data fusion method based on the Gaussian process prior to infer the customer-choice-occasion specific preferences in the fused experimental-database data.

3.3.3 Gaussian Process Prior

Gaussian processes are continuous stochastic processes and can be considered as a probability distribution over random functionals, generalizing the multivariate normal distribution (over random variables). The Gaussian process prior is used in statistical modeling and machine learning for nonlinear regression models due to the wide range of functional forms that it can capture. As a stochastic process, the parameter set (or indices) of the Gaussian process may be unidimensional linearly ordered sets such as *time*, or more general mathematical sets such a D-dimensional continuous *space*. In the pioneering marketing application of GPP, Dew and Ansari (2017) employed a series of unidimensionally *time*-indexed Gaussian processes to capture nonlinear time-trends in consumer purchase behavior. In the present study, we utilize the GPP to model α_{ijt} as realizations on a random functional over the *space* that spans the customers (i) and their subscription choice (j) occasions (t). We denote the *parameter set* that indexes this customer-choice-occasion space as,

$$z_{111}, \dots, z_{ijt} \dots, z_{NJT} \in \mathbb{R}^D$$

where z is a D-length vector of input features (e.g., demographic information, prices, calendar and membership times) corresponding to observation ijt. Given inputs z, the GPP of α_{ijt} can be fully characterized by a mean function $m(z)$ and a covariance function, or kernel, $k(z, z')$:

$$\alpha_{ijt} \sim \mathcal{GP}\big(m(z), k(z, z')\big) \qquad [3.4]$$

where z' denotes the set of input features of any other customer-choice-occasion observations. As α_{ijt} is the choice model intercept term, we specify the mean function to be the static alternative-specific intercept α_j. Moreover, letting

$$\mathrm{A}_j := \left\{\left\{\alpha_{ijt}\right\}_{i=1}^{N}\right\}_{t=1}^{T}$$

we exploit the property that for any finite set of customers and their choice occasions, the marginal of the Gaussian process functional values is distributed multivariate normal:

$$A_j \sim MVN\left(\boldsymbol{\alpha}_{\mathbf{j}}, K(\mathbf{z}, \mathbf{z}')\right) \qquad [3.5]$$

where K is the NT-by-NT covariance matrix with elements consisting of $K_{i,i'} = k(z, z')$.

3.3.4 GPP for Data Fusion

As a Bayesian nonparametric prior, Gaussian processes enable shrinkage over nonlinear random functionals. Specifically, the kernel function $k(z, z')$ provides a measure of similarity between observations and regularizes the degree to which any realization of α_{ijt} can deviate from the mean function (α_j) relative to any other $\alpha_{i'jt}$. We will make use of a kernel that is robust to the inclusion (exclusion) of specific time-varying covariates, which we formally introduce in the next section. It is this kernel mechanism that our data fusion framework enables inference on the one-shot experimental subjects' long-run subscription choices despite the inherent missingness of this information. The longitudinal renewal patterns of the database customers regularize, or *inform*, the likely trajectory of the experimental subjects' future behavior based on the similarity of their corresponding observables among $\boldsymbol{z}$. In other words, GPP serves as a fusion mechanism that may *most closely thought of* as matching the experimental subjects whose renewal choices we do not observe, to their closest counterparts in the longitudinal database, for whom renewal outcomes are observed.

However, in comparison to existing matching-based fusion techniques, there are several advantageous to using GPP. First, there is no explicit matching step, which Gilula et al. (2006) argue are often specified *ad hoc* and sensitive to parametric, covariate, and linearity restrictions (e.g., propensity score matching). Rather than explicitly imputing a match and estimating the model, GPP works by enabling between-customer and between-choice-occasion shrinkage of the

baseline preference (α_{ijt}), which is integrated into the model likelihood. Next, although stratification techniques have been developed to alleviate the one-to-one outcome of standard matching methods, (1) these within-strata matches remain equally weighted, (2) cutoffs are difficult to assign, and (3) observations outside the stratum are *ex ante* omitted from informing the matched record. The latter issue could be problematic on large CRM databases where non-stratum observations can be individually of negligible informativeness, yet in aggregate highly informative. In comparison, our proposed GPP-based fusion enables the sharing of information across all customers using in-model regularization, that differentially modulate across people and choice occasions. Lastly, while multiple imputation matching can overcome the lack of uncertainty propagation from imputation to estimation (Andridge and Little 2010, Feit and Bradlow 2016), it often comes at the cost of assuming CIA for these methods to remain computationally tractable (cf. Qian and Xie 2014). We obviate this concern as the GPP covariance function is fully-specified and captures dependencies (e.g., similarity) between all observations across both datasets, and we provide an efficient estimation strategy using stochastic variational inference to scale the GPP fusion framework to large data.

Taken together in the context of the fused experimental-database data, we employ a Gaussian process prior as the mechanism to allow for and regularize the sharing of information over the joint *time-space* of customers and subscription occasions to enable inference on individual-level dynamic preference heterogeneity. Reproducing our focal model (Eqs. 3.3a and 3.3b) and generalizing to the fused data:

$$F^{-1}(Y_{it} = j \mid p_{ijt}, X_{it}, \{X_{i'}\}, \Theta) = \alpha_{ijt} + p_{ijt}\beta + X_{it}\gamma_j + \mu\varepsilon_{ijt} \qquad [3.6a]$$

$$\alpha_{ijt} \sim \mathcal{GP}\left(\alpha_j, k_{ARD}\,(z, z')\right) \qquad [3.6b]$$

where $z \coloneqq \left(X_{it}, p_{ijt}, t_i\right)$ and z' denote any set of corresponding input features belonging to the other customers $\{X_{i'}\}$. In line with the general form (Eq. 3.5), the data fusion GPP (Eq. 3.6b) is fully characterized by its mean function, given here by the static alternative-specific intercept α_j, and its covariance function, which

we specify using the Radial Basis Function Automatic Relevance Detection (ARD) kernel (Neal 1996) over the D-dimensional tuple (z, z'):

$$k_{ARD}(z, z') = \eta^2 \exp\left(-\frac{1}{2}\sum_{d=1} \frac{\|z_d - z_d'\|^2}{\rho_d^2}\right) \quad [3.7]$$

3.3.5 Automatic Relevance Detection (ARD) Kernel

The ARD kernel is the multidimensional generalization of the ubiquitous squared-exponential (SE) kernel, the latter of which is unidimensionally indexed. To reiterate, rather than a unidimensional index set such as time, the proposed data fusion GPP is indexed over the space of customer-choice-occasions whose parameter set consist of the observables given by $z \coloneqq (X_{it}, p_{ijt}, t_i)$, which is the joint space of the customers' characteristics, price offers, and time trends. The choice of the ARD kernel is by the GPP's role within our fusion framework to regularize the sharing of information across customers-choice-occasions with respect to the similarity of their corresponding z.

Whereas the unidimensional SE kernel is characterized by two hyperparameters (amplitude and lengthscale), there are $D + 1$ ARD hyperparameters: the amplitude of the Gaussian process η^2, and D feature-specific lengthscales ρ_d^2. As the amplitude of the covariance function based on the ARD kernel is governed by a single amplitude (η^2), the individual lengthscale hyperparameter (ρ_d^2) represents the standardized influence of different input features $\{z_d\}$ on the functional outputs. As such, the "automatic relevance detection" capacity of the kernel arises by shrinking ρ_d^2 to zero for features that do not inform the evolution of α_{ijt} across customer-choice-occasions, and by extension, the fused data. Specifically, this mechanism strengthens the robustness of the GPP data fusion to the inclusion (exclusion) of specific time-varying covariates.

Shown in Fig. 3.5 across two dimensions, the ARD kernel (Eq. 3.7) is equivalent to the product of D unidimensional SE kernels, $k_{ARD}(z, z') = \prod_{d=1}^{D} k_{SE}(z_d, z_d')$. Here, the nonlinearities captured by the separate Gaussian processes using SE kernels over the individual features are joined together in the single GP with an ARD kernel over the *space* of both features. Recalling that the linear parameters (β, γ_j) of the data fusion model (Eq. 3.6a) capture only the main-

effects of the covariates, this attribute of the ARD kernel is central to the model's ability to capture any nonlinear evolution of a_{ijt} across customer-choice-occasions, as governed its GPP over the *space* of z.

Taking the perspective that the marginal distribution of a Gaussian process over a finite set of observation is MVN (Eq. 3.5), it follows that specific realizations of a_{ijt} are correlated random-variates through the covariance matrix $K(\mathbf{z}, \mathbf{z}')$, which in our data fusion framework has $K_{i,i'} = k_{ARD}(z, z')$. Therefore, within the marginal distribution of the observed fused data, the sharing of information between elements of $\{a_{ijt}\}$ is governed by the correlations in the entries of the covariance matrix. For the ARD kernel, these correlations are a function of the Euclidean norm on the input features, $\|z_d - z'_d\|$. In other words, given any pair of customer-choice-occasion observations, as the input features (z, z') that characterize them begin to differ, their respective a_{ijt} terms are less correlated and thereby less likely to be informative onto one another. Hence, for the purpose of augmenting the experimental subjects with a renewal trajectory, the GPP fusion can then make use of all available customer-choice-occasion information from the database. The ARD kernel flexibly determines the degree to which the time-varying preferences of any user from the database will inform the long-run trajectory of a subject from the experiment.

3.3.6 Augmenting the One-Shot Experiment

We now turn our attention to addressing the initial stated task of augmenting the one-shot experiment with long-run renewal trajectories at the individual subject-level, by borrowing information across the fused data. This problem is equivalent to drawing a forward-period functional value from the Gaussian process posterior predictive distribution, which conveniently exists in closed-form. Let A_j (Eq. 3.5) denote the vector of intercepts estimated for *observed* choices in the combined data, which consists of the initial upgrades of the one-shot experiment and all subscription-choice observations from the longitudinal database. Conditional on A_j, $\boldsymbol{z}$, and the kernel hyperparameters, the forward-prediction (a_{ijt^*}) of experimental subject i for a future renewal occasion $t^* > 1$ is given by:

$$\begin{bmatrix} A_j \\ a_{ijt^*} \end{bmatrix} \sim MVN\left(\begin{bmatrix} a_j \\ \vdots \\ a_j \end{bmatrix}, \begin{bmatrix} K_{ARD}(\boldsymbol{z}, \boldsymbol{z}) & K_{ARD}(\boldsymbol{z}, z^*) \\ K_{ARD}(z^*, \boldsymbol{z}) & K_{ARD}(z^*, z^*) \end{bmatrix} \right)$$

[3.8]

By property of the MVN, the univariate normal conditional distribution of the forward-prediction is:

$$\alpha_{ijt^*} \sim N\left(\mu^*, {\sigma^2}^*\right) \quad [3.9]$$

Where,

$$\mu^* = a_j + K_{z^*,\boldsymbol{z}} K_{\boldsymbol{z},\boldsymbol{z}}^{-1}\left[A_j - \alpha_j\right]$$
$${\sigma^2}^* = K_{z^*,z^*} - K_{z^*,\boldsymbol{z}} K_{\boldsymbol{z},\boldsymbol{z}}^{-1} K_{\boldsymbol{z},z^*}$$

Note that while subject i's initial subscription ($t = 1$) characteristics are elements within $\boldsymbol{z}$, the forward-prediction inputs z^* are not wholly identical to these due to time-varying features, which must be accordingly extrapolated (e.g., membership duration, past subscription choices). Having drawn a forward prediction α_{ijt^*}, it can be applied to calculate $F^{-1}(Y_{it^*} = j \mid p_{ijt^*}, X_{it^*}, \{X_{i\prime}\}, \Theta)$ to generate predictions on $\Pr(Y_{it^*} = j)$. These are the database augmented choice probabilities needed to estimate the long-run CLV of the experimental subject (Eq. 3.1).

3.4 DATA DESCRIPTION

For the empirical application of the proposed data fusion framework, we obtained novel large-scale customer-transaction-level data from a leading U.S.-based online dating site, composed of the firm's complete panel-structured CRM database as well as an orthogonalized subscription pricing experiment on first-time users. The experiment was designed by the authors of this study in partnership with the company, conducted on a randomly allocated sample of first-time users across a one-month window. We refer to this experiment as 'one-shot' given its cross-sectional format due to having outcome observations solely on the initial choices of the new-user sample. In fusing these two datasets, our goal is to draw upon the longitudinal variation of the CRM database to infer the long-run CLV trajectory of the experimental participants.

Among both datasets, the customer lifecycle begins with their initial registration onto the site as a free user. All interactions with the site are preceded

by a login, which we define as the primary activity-level covariate (LOGINS). Free users may choose to become a paid member by purchasing a 1-, 3-, or 6- month contract, which affords the customer the ability to initiate contact with any other users, whereas free users can only respond to initiated contacts. At the end of a contract period, users are presented with choice to renew their contract or switch to an alternate option, including reverting to a free user. As such, the multinomial subscription choice set consists of four options: (No Subscription, 1-Month, 3-Month, 6-Month).

3.4.1 Longitudinal CRM Database

The CRM database is composed of the site's user profiles, partner-seeking preferences, purchase transactions, price offers, and login information between January – October 2013. Over this period, 530,291 first-time users joined the site, for whom we are able to observe their complete purchase and activity histories and refer to as *in-period users*. In addition, there are 249,813 'active' *existing users*, defined as those who joined prior to January 1 meeting the criterion of either (1) being under an active subscription contract, or (2) for non-subscription periods, having made at least one login within a 30-day window. We observe their in-period purchase, as well as the length of their membership since joining (DAYSJOIN) and days since last paid subscription (DAYLASTSUB). The earliest join date of active existing users is from July 2001. Taken together, the CRM database consists of purchase incidences between January and October 2013 of customers who joined between July 2001 and October 2013.

The profile data consist of 50 types of variables on demographics, free text (e.g., bio and other self-descriptions), location, partner-seeking preferences (e.g., distance, age and height range), and a proprietary 'relationship style' categorical variable that we rename as MATCHTYPE (to prevent revealing the site's identity). A crucial consequence of the site's algorithm for MATCHTYPE is that all profile variables are required in building the initial user profile[9], with exception of the free texts. We also observe profile changes over time, with partner-seeking preferences exhibiting most of the temporal variation, and demographics the least.

[9] In generating their public profiles, users have a high degree of flexibility to obfuscate any of the mandatory profile information. We do not consider data related to this feature, as from the firm's perspective, all profile variables are observed.

Our model's outcome of interest is the panel of user-specific purchase incidences with respect to the four contract choices. There are ~1.3 million purchase incidences within the longitudinal database that we map to corresponding price offers that the customers are presented to at the time. At any purchase decision point, a customer has available only one set of price offers to choose from. We exclude any transactions that are cancelled by the user, as the site allows a 24-hour grace period for any subscription purchase. Moreover, to account for potential lagged effects of choice and pricing, we incorporate prices and choice outcome from the most recent prior transaction to each transaction records. In the case of initial purchases by new-users, the lagged transaction information is defaulted to 'No Subscription', and prior prices equaled to current price offers. Pricing structure and choice patterns of the site are discussed in more detail in the next section.

We also join onto each purchase transaction, the recency and frequency of logins that occurred in the period between the current and the most recent past transactions. Lastly, in forming the complete purchase incidence observation (Y_{ijt}, X_{ijt}), profile and preference information are mapped by the most recently updated information prior to the purchase timestamp.

3.4.2 One-Shot Pricing Experiment

The experiment was conducted on a sample of 18,286 first-time users who joined the site in February 2014 and were randomly assigned to one of twenty pricing conditions designed to orthogonalize across the 1-, 3-, 6- month options. This design was meant to address a common pathology to the pricing regimes observed on the site, and typical to subscription services, whereby all contract prices are modulated contemporaneously, e.g., an 'across the board' price cut of some percentage. Our orthogonalized design decorrelates the variation between the contract prices *across* the treatment conditions. Throughout the trial period, the test participants are exposed only to pricing associated with their treatment. They are unaffected by any concurrent pricing regimes the site may offer, and the site does not list prices publicly. Moreover, the experimental price ranges exceed at a minimal the 99% interval of all price offers observed in the database centered to the medians. The experimental price range of the 1-month option is between $24.99 and $36.49 (vs. $26.99 to $31.99 in 99% interval centered around the database 1-month median), the 3-month between $53.97 and $65.97 (total, vs. $56.97 to $63.97

in the database 99% w.r.t. median), and 6-month between $71.94 and $101.94 (total, vs. $75.33 to $83.94). It should be noted that the experiment's maximum price point in all three contracts exceed those observed in the database.

For each test subject, we observe their activities and purchases on the site for an additional four weeks. However, this window precludes the ability to observe any subsequent renewals as the minimal contract length is 1-month, a timeframe that would exceed the closest renewal opportunities even for users who had signed up at the very beginning of their four-week window. Although it is possible that first-time users may make their purchase decision after the four-week period, evidence from the site's CRM database suggest that the vast majority users who ever make an initial purchase do so well within this period of time. We detail this below, along with additional pricing and purchase incidence descriptives.

3.4.3 Purchase and Renewal Incidence: Model-Free Descriptives

For users who ultimate upgrade to a paid subscription, 89.4% of their initial purchases occur within the first 4-weeks of registration. Although the distribution of time-to-initial-purchases exhibits a long right-tail (Fig. 3.1), users who do make their initial purchases beyond the first month are typically responding pricing offers different than those introduced during their initial sign-up. As such, we consider these customers to have been nonresponsive with respect to the initial price offers. Moreover, we observe that in both the database and experiment a precipitous drop in initial purchase incidences, such that for users who sign-up on the same date, the number of those who upgrade on the 28th day is only 2.9% of the number of those who upgrades on their 1st day on the site. We account for the potential effect of membership duration on purchase propensity by including time-trend covariates for days since registration (DAYSJOIN) and days since last paid subscription (DAYLASTSUB), both of which also enter into the ARD kernel of the Gaussian process prior.

Across the calendar year, the number of total incidences of *initial purchases* are relatively stable across weeks. Notable exceptions for the site are the weeks of Valentine's Day, Easter, Memorial Day, and Labor Day when purchase incidences rise (Fig. 3.2). This is likely in part due to price discount promotions that the website runs annual during these periods, as well as the increased interest in seeking romantic partnerships during long weekends and holidays (e.g., Valentine's

Day being a ubiquitous example). Rather than explicitly defining holiday and weekly fixed effect terms, which may overlook or obfuscate time-trends across the calendar year that do not adhere to common time denominations, our approach to controlling for calendar-time trends follows Dew and Ansari (2018). Among the features that index the ARD-GPP is a continuous DAYOFYEAR term that serves to flexibly capture calendar-time fluctuations in choice propensities. Notably, although we additionally include this term as a linear covariate, we show that the linear (main) effect is non-significant. This is intuitively in line with the evidence from Fig. 3.1, where there is no distinguishable directional trajectory of purchase incidences over the calendar year beyond seasonality and holiday time-trends that would be readily captured by the GP.

Among first-time users, 34.1% eventually purchase at least one paid subscription during the customer lifecycle (Fig 3.3). Interestingly, their preferences among the three subscription contracts are not ordinal to the contract lengths, but rather consists of the 1-month followed by the 6- and 3- months, respectively. This may suggest that most customers view membership duration dichotomously, i.e., they'll either be a short- or long- term customer, whereas the 3-month option potentially serves a tertiary group of those who fall in-between these two larger clusters. Across choice occasions, we find that customers are unlikely to switch into other contract options, except when in choosing the 'No Subscription' option to end their paid membership (Table 3.1a). Overall, the number of customers who renew their subscriptions drop-off at a rate that is roughly inverse-proportional to the number of renewal occasions (Fig. 3.4). Some interesting trends emerge when breaking down the contract renewal transition matrix by the first three renewal occasions. As before, we find that churn rates across subscription lengths remain relatively stable, and most subscribers who do renew their subscriptions, remain in choosing their previous contract. Interestingly, 6-month subscribers begin to transition into the 1-month contract across renewal occasions. The 3-month subscribers who transition to other contracts are almost evenly split between 1- and 6- months, whereas 1-month subscribers almost never move into a longer-duration contract.

Taken together, the model-free evidence on longitudinal subscription patterns suggests the importance in understanding the impact of initial pricing on long-run CLV on this site. In particular, if first-time users are to become paying subscribers,

they are most likely to do so within early on. Second, the contract choice of the initial purchase can strongly influence subsequent renewal choices, churn rate, and consequently, customer lifetime value. While the site can naively wait out an intended period of time to measure the long-run effects of their initial manipulations, or worse myopically undertake optimization based on the one-shot outcomes, in fact much of the desired variation in consequent behavior already exists in the CRM database. In the next section, we formally develop the Bayesian nonparametric data fusion framework that will allow the site to augment the one-shot experiment with database information to make expedited long-run inferences.

3.5 ESTIMATION

Having introduced the utility structure of the Gaussian process prior data fusion framework for multinomial subscription outcomes (Eqs. 3.6a and 3.6b), we now derive the Bayesian posterior inference strategy for two widely used specifications for polytomous choice settings: the multinomial logit (McFadden 1973) and probit (Albert and Chib 1993) models. As our fused experiment-CRM dataset involves large-scale choice data on user-level transactions (1.3 million) that would otherwise be infeasible using standard MCMC, we employ stochastic variational inference (Hoffman et al. 2013) to transform the high-dimensional integration problem as a tractable optimization problem.

3.5.1 Stochastic Variational Inference (SVI)

The advantage of variational Bayesian estimation lies in its potential for vast scalability in empirical applications, in terms of both computing time and resource, where posterior inference is sought. Variational Bayes provides a principled framework to approximate the true posterior by *maximizing* the accuracy of its estimates through *minimizing* the Kullbeck-Leibler (KL) divergence, a similarity measure, between the posterior density $p(\theta|x)$ and an approximating variational density $q(\theta)$, which is otherwise equivalent to optimizing the evidence lower bound (ELBO) on the marginal likelihood (Jordan et al. 1999). The choice of the approximating distribution $q(\theta)$ is of central importance in formulating variational objective functions, impacting the degree to which KL divergence can be in-principle minimized, the accuracy of the estimates, and the scalability at runtime.

A key advantage to variational approximation is that the complete data likelihood is unaltered, but rather augmented with "approximating" densities to jointly form a single optimization bound. The simplest and most common choice of $q(\theta)$ is to assume a fully factorized distribution, $q(\theta) = \prod_{\ell}^{L} q_{\ell}(\theta_{\ell})$, referred to as mean-field variational inference. An advantage of the mean-field independence approximation lies in its ease of implementation, including fully automatically by black-box software and auto-differentiation optimization packages such as Stan. However, this simplification comes at a trade-off against the fidelity of the posterior approximation and can result in local optima (Hoffman and Blei 2015). It is therefore at least partially due to the difficulty in selecting the appropriate form of $\{q_{\ell}(\theta_{\ell})\}$ along with the uncertainty in posterior convergence that have resulted in noticeably sparse applications of variational inference by marketers (Braun and McAuliffe 2010, Dzyabura and Hauser 2011, Puranam et al. 2017, Ansari et al. 2018), despite the technique's potential for both accurate and scalable inference on posteriors, as well as prevalent application in other domains including natural language processing, computer vision, and collaborative filtering.

In order to apply the proposed data fusion framework onto the fused experiment-CRM dataset, we derive a set of novel *stochastic variational inference* (SVI) algorithms for polytomous choice models with Gaussian process priors, based on the multinomial logit (MNL) and probit (MNP) specifications. The key motivation of our estimation approach is the need to overcome the cubic computational complexity of Gaussian processes, $\boldsymbol{\mathcal{O}}(n^3)$, that renders the models inestimable on our dataset. More generally, the scale of our empirical context increasingly represents the norm than the exception for most marketers. As a result, classical estimation methods for GP, including MCMC, have thus far limited its application in modern database marketing problems. To this end, the estimation strategy we develop here represents a broader contribution of incorporating the flexibility of Gaussian process priors to large-scale marketing response models.

Among variational Bayes, SVI improves over mean-field variational inference on several criteria as it allows for: correlation structures between relevant subsets of parameters (fidelity), principled part-wise update on minibatches of data (scalability), and in our specific derivations, analytic solutions for natural gradients on the ELBO that provide second-order optimization updates (efficiency). In

deriving the SVI algorithms for Gaussian process prior MNL and MNP data fusion models, we draw upon several lines of interrelated Bayesian and machine learning literature: sparse 'inducing point' Gaussian processes (Snelson and Ghahramani 2006, Titsias 2009, Hensman et al. 2015), efficient ELBO formulation to polytomous outcomes (Titsias 2016, Ruiz et al. 2018), Pólya-Gamma data augmentation for logistic regressions (Polson et al. 2013, Wenzel et al. 2018), and exact Hamiltonian Monte Carlo quadrature for probit utilities (Pakman and Paninski 2012). Specifically, our innovation is twofold: generalizing the SVI framework for Pólya-Gamma augmented sparse GP *binary* logit (Wenzel et al. 2018) to a multinomial setting, and extending the stochastic variational expectation-maximization (SVEM) strategy for multinomial probit (Ruiz et al. 2018) to use a more efficient exact HMC E-step. We provide these variational bounds (detailed derivations in Appendix B).

3.5.2 Sparse Gaussian Process

At the heart of our scalable inference strategy for both model specifications is the sparse Gaussian process estimation (Titsias 2009). This approach reduces the cubic computational complexity $\mathcal{O}(n^3)$ of direct inference on GP models to $\mathcal{O}(nm^2)$, where $m \ll n$ and indicates the dimensionality of a set of variational auxiliary parameters known as 'inducing points' augmented onto the ELBO. At the heart of this technique's computational efficiency is the avoidance of direct inversions on the n-by-n covariance function of GPs when calculating their marginal density. Snelson and Ghahramani (2006) first suggested the use of a set of m points within the same space of the n observations that could reproduce the functional shape of the corresponding GP. Intuitively, as it's likely that datasets with large n can contain observations with redundant informativeness to certain ranges of the GP functional, they can be replaced by a more succinct set of 'inducing points' with negligible loss of accuracy. However, as the shape of the GP functional is unknown *a priori*, the optimal location of these inducing points is also unknown and may not need to correspond to actual data points. From a variational perspective, these inducing points can then be viewed as latent variables to be estimated as part of the ELBO.

The derivation of the optimal variational density for the inducing point is extensive and has received much attention in the Bayesian and machine learning

literature (cf. Gal and van der Wilk 2014 for a detailed technical report). Here, we reproduce the key identities necessary to incorporate sparse GP into the ELBO of our Gaussian process data fusion framework. Decomposing the deterministic component of utility, $\boldsymbol{\psi} = \boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}\boldsymbol{\beta}$ is the linear component and $\boldsymbol{\alpha}$ is the vector of $\{a_{ijt}\}$ having the ARD-kernel Gaussian process prior (Eq. 3.6a and 3.6b), we augment the latent GP functional over $\boldsymbol{\alpha}$ with a set of m additional 'inducing point' auxiliary variables $\boldsymbol{u} = \{u_1, \ldots, u_m\}$ on the functional *outputs* along with a set of corresponding 'inducing points' on the covariate *inputs* (latter notation suppressed). The optimal augmented distribution of $\boldsymbol{\alpha}$ and variational distribution of $\mathbf{u}$ are then given by,

$$p(\boldsymbol{\alpha}|\boldsymbol{u}) = MVN\left(\boldsymbol{\alpha} \mid K_{nm}K_{mm}^{-1}\boldsymbol{u},\, \widetilde{K}\right)$$

$$p(\boldsymbol{u}) = MVN(\boldsymbol{u} \mid 0,\, K_{mm})$$

where K_{mm} in our implementation is the ARD kernel covariance function formed over the input inducing points, K_{nm} is the nonsymmetric rectangular 'cross-kernel' matrix between observed data and inducing points, and $\widetilde{K} = K_{nn} - K_{nm}K_{mm}^{-1}K_{mn}$. It's worth noting that K takes an analogous form to the conditional multivariate Gaussian covariance term, reinforcing the relationship in $p(\boldsymbol{\alpha}|\boldsymbol{u})$.

Lastly, the choice of the quantity m represents a tradeoff between accuracy of capturing the functional and computational speed. Hensman et al. (2015) shows that when $m = n$, the optimal sparse GP augmented density (Eq. B10) reverts to the standard Gaussian process specification. Typically, m needs only to be a small proportion of n to fully capture the GP, as the inducing points need only be optimally placed at locations where the functional shifts from concave to convex, and vice versa, including saddle points in the multidimensional space of the ARD kernel's index set. For example, dozens of inducing points would already have the ability to capture up to what can be considered as a complex GP functional shape. Moreover, a key feature of the sparse GP when taking a variational approach, a trace term of $\widetilde{K}$ arises in the ELBO and serves as an *overfitting penalization* factor when the predefined m exceeds the number of inducing points necessary to recover

the functional shape (Titsias 2009). This penalty term results in excess inducing points to closely cluster around functional ranges with high certainty, so as to avoid placements elsewhere. Ultimately, the choice of m is dependent on the empirical context, and we demonstrate the accuracy of recovery across a range of m in simulation studies.

3.5.3 Multinomial Logit

In forming the ELBO for the MNL, we generalize the GPP binary logit framework of Wenzel et al. (2018) using the one-vs-all lower-bound form of Titsias (2016) to maintain analytical tractability. Wenzel et al. (2018) builds upon the Pólya-Gamma data augmentation strategy of Polson et al. (2013) that admits closed-form updates for Bayesian inference on the MNL, including MCMC and variational Bayes. We provide details on the derivations of the complete MNL lower-bound in Appendix B, and reproduce the key identities here.

The one-vs-all MNL lower bound (Titsias 2016) is given by rewritten the multinomial softmax function as,

$$p(y_{ijt} = j|\psi) = \frac{1}{1 + \sum_{j' \neq j}^{J} \exp\left(-(\psi_{ijt} - \psi_{ij't})\right)} \qquad [3.10]$$

This can be bounded by replacing the summation in the denominator of Eq. 3.10 with a product over sigmoid functions,

$$p(y_{ijt} = j|\psi) \geq \prod_{j' \neq j} \frac{1}{1 + \exp\left(-(\psi_{ijt} - \psi_{ij't})\right)} \qquad [3.11]$$

$$= \prod_{j' \neq j} \sigma(\psi_{ijt} - \psi_{ij't}) \triangleq \boldsymbol{\mathcal{L}_1}$$

where $\sigma(\cdot)$ is the sigmoid function and the inequality holds because $(1 + \sum_i n_i) \leq \prod_i (1 + n_i)$ for all positive real numbers. Critically, Titsias (2016) shows the optimality of this bound as maximizing the true likelihood based the multinomial softmax function exacts the same parameter estimates as optimizing over the exact data likelihood, which we denote as $\boldsymbol{\mathcal{L}_0}$. This result holds because $\boldsymbol{\mathcal{L}_1}$ is a concave function of $\{\psi_{ijt}\}$, such that the stationary conditions when maximizing over $\boldsymbol{\mathcal{L}_0}$ also globally maximizes $\boldsymbol{\mathcal{L}_1}$. The computational advantage of

this bound lies in its factorization of the standard softmax into a series of products that each only depend on a pair of utility values $(\psi_{ijt},\ \psi_{ij't})$. This now allows us to proceed with forming the second bound on the GPP-MNL specification utilizing the Pólya-Gamma data augmentation strategy (Wenzel et al. 2018), crucially by preserving its single-bound closed-form solution.

Polson et al. (2013) posited the errors of binomial distributions, including Extreme Value type distributions, as a scale mixture of Gaussians under the Pólya-Gamma (PG) distribution. At the heart of the PG distribution is the equivalence of a generalized sigmoid function to the expectation over the PG random-variate (ω) scaled quadratic exponential (e.g., Gaussian) term,

$$\frac{(\exp(\psi))^a}{(1+\exp(\psi))^b} = 2^{-b}\int_0^\infty \exp\left(\kappa\psi - \frac{\omega\psi^2}{2}\right)p(\omega)d\omega \qquad [3.12]$$

For logistic regressions[10], $a, b = 1$, $\kappa = y_j - 1/2$ for $y_j \in \{-1, 1\}$ denoting whether the choice outcome was alternative j, and $\omega \sim PG(1, 0)$. When applying the one-vs-each $\boldsymbol{\mathcal{L}_1}$ (Eq. 3.11) for the softmax to proceed in generalizing Wenzel et al.'s (2018) approach to MNL, this results in:

$$\begin{aligned}\boldsymbol{\mathcal{L}_1} &= \prod_{j'\neq j}\sigma(\psi_{ijt} - \psi_{ij't}) \\ &= \prod_{j'\neq j}\frac{1}{2}\int_0^\infty \exp\left(\frac{(\psi_{ijt} - \psi_{ij't})}{2} - \frac{(\psi_{ijt} - \psi_{ij't})^2}{2}\omega_{ijt}\right)p(\omega)d\omega\end{aligned} \qquad [3.13]$$

To formulate the final variational bound on the MNL specification of the GPP data fusion framework, we apply the Jensen's inequality on the conditional log-likelihood of the observed choice outcomes, and introduce the variational distributions of the augmented PG variates $q(\boldsymbol{\omega})$ as well as the sparse GP inducing points $q(\boldsymbol{u})$ on this inequality on the likelihood marginalizing over all parameters results in the ELBO,

$$\begin{aligned}\log p(\boldsymbol{y}) \geq\ & \mathbb{E}_{p(\boldsymbol{a}|\boldsymbol{u})q(\boldsymbol{u})q(\boldsymbol{\omega})}[\log p(\boldsymbol{y}|\boldsymbol{\psi}, \boldsymbol{\omega})] - KL(q(\boldsymbol{u})||p(\boldsymbol{u})) \\ & - KL(q(\boldsymbol{\omega})||p(\boldsymbol{\omega}))\end{aligned} \qquad [3.14]$$

[10] See Polson et al. (2013) section 3 for proof.

where $q(\omega_{ijt}) = PG(\omega_{ijt} \mid 1, c_{ijt})$ and $q(\boldsymbol{u}) = MVN(\boldsymbol{u} \mid \boldsymbol{\mu_u}, \Sigma_{\mathrm{u}})$. Note that the variational distributions for $\boldsymbol{\omega}$ and $\boldsymbol{u}$ introduce an additional layer of local (observation-level) parameters $\{c_{ijt}\}$ and global parameters $\boldsymbol{\mu_u}, \Sigma_{\mathrm{u}}$ that now enter into the ELBO. These parameters enable the decomposition of the optimization updates of the ELBO objective function across minibatches data to allow for scalable stochastic variational inference (Hoffman et al. 2013, Wenzel et al. 2018). The final analytical form of the complete ELBO for the GPP-MNL is given below (reproducing Eq. B14-B19). This form is amenable to efficient and scalable natural gradient-based updates over minibatches of data,

$$\begin{aligned}\boldsymbol{\mathcal{L}}(\boldsymbol{c}, \{\boldsymbol{\mu}_u\}, \{\boldsymbol{\Sigma}_u\}) = \frac{1}{2}\Bigg(& \sum_{j \neq j'} \Big(\widetilde{\Upsilon} - \widetilde{\Upsilon}^{\mathrm{T}} \Xi \widetilde{\Upsilon} - tr\big(\Xi \widetilde{\mathrm{K}}_j + \Xi \widetilde{\mathrm{K}}_{j'}\big) \\ & - tr\left(\boldsymbol{\kappa}_j^{\mathrm{T}} \Xi \boldsymbol{\kappa}_j \Sigma_{uj} + \boldsymbol{\kappa}_{j'}^{\mathrm{T}} \Xi \boldsymbol{\kappa}_{j'} \Sigma_{uj'}\right)\Big) \\ & - \sum_{J}^{I,J,T} 2 \log\cosh\left(\frac{c_{ijt}}{2}\right) - \frac{c_{ijt}}{2} \tanh\left(\frac{c_{ijt}}{2}\right) \\ & - \sum_{j=1} tr\big(K_{jmm}^{-1} \Sigma_{\mathrm{j}}\big) + \boldsymbol{\mu}_{uj} K_{jmm}^{-1} \\ & - \log\left|\Sigma_{uj}\right| + \log|K_{mm}| \Bigg)\end{aligned} \qquad [3.15]$$

where,

$$\Xi = diag\big(\{\xi_{ijt}\}\big), \xi_{ijt} = \mathbb{E}_{p(\omega_i)}[\omega_{ijt}] = \frac{1}{2c_{ijt}} \tanh\left(\frac{c_{ijt}}{2}\right)$$

$$\widetilde{\Upsilon} = \mathbb{E}_{p(\boldsymbol{u})}[\Upsilon] = \big(K_{jnm} K_{jmm}^{-1} \boldsymbol{\mu_{uj}} + X\beta_j\big) - \big(K_{j'nm} K_{j'mm}^{-1} \boldsymbol{\mu_{uj'}} + X\beta_j\big)$$

$$\boldsymbol{\kappa}_j = K_{jnm} K_{jmm}^{-1}$$

3.6 RESULTS AND PLANNED ANALYSIS

We now turn to the results and planned direction of the empirical application of the Bayesian nonparametric data fusion framework. We estimated the GPP-MNL specification of the framework using the stochastic variational inference strategy described, fusing experimental participants who've made a subscription

purchase during their 4-week trial period ($N_{sub}^{E} = 3{,}758$) to the full longitudinal CRM database ($N_{all}^{L} = 780{,}104$). Experimental participants who did not subscribe during the trial period are omitted from the analysis for two reasons: (1) our metric-of-interest is CLV until first 'No Subscription' outcome, and (2) given that we wish to infer initial pricing effects on the CLV metric, customers who do not respond to the experimental pricing conditions are considered to have not responded their assigned initial pricing. The decision to measure CLV until the first non-subscription choice reflects that nearly all users who unsubscribe from their paid subscriptions do not return to the site, an attribute of customer lifecycles typical to online dating services focused on partner-seeking for marriage and serious relationships, as is the case here. Moreover, as experimental participants' assigned price conditions expire upon the conclusion of the 4-week trial period, we consider those who do not subscribe as nonresponsive to their initial pricing, e.g., choosing the default 'No Subscription' with respect to their price treatment.

3.6.1 CLV Holdout Sample

To demonstrate the predictive accuracy of our data fusion CLV model, we utilized a holdout sample of new users ($N_{out} = 5{,}220$) who joined the site in the same weekend of January 2013 and were selected as part of a "one-shot" initial pricing A/B test. These customers were randomly assigned to one of eight pricing conditions arising from a 2-by-2-by-2 design: high-vs-low price for each of three subscription contracts.

As shown in Table 3.1, all price conditions for the holdout sample are at or below the site's standard prices. The A/B test ran for a period of 7 days, and as such, is comparatively limited in terms of both price variations and trial length with respect to the experiment designed by the authors of this study. However, as the experiment coincides with the start of the CRM observation window, we are able to observe all purchase incidences of the holdout sample between January to October 2013, and thereby observe their in-period CLV.

To compare the data fusion framework, the benchmark models that we utilized are propensity score matching (Rubin 1973), MNL model fitted on the cross-sectional variation of the experiment only, and a random-effect MNL with time-trend fixed-effects on the panel variation of the CRM database. We provide

outcomes in terms of mean absolute percentage error (MAPE) for each method's estimated CLV (Eq. 3.1)

In Table 3.2, we find that the data fusion framework to exhibit the lowest CLV MAPE followed by the CRM random-effect, propensity score matching, and lastly, the experiment-only model. The poor fit of the experiment-only model can be attributed to the model's lack in both model parameters and data variation that capture temporal variation in subscription choices. While propensity score matching improves over the experiment-only model as we allow for time-varying observables (renewal occasion, membership length, and past subscription prices and choices), the method's fit is likely limited by a lack of terms that capture preference heterogeneity across customers. The CRM-only model incorporates both time-trends and preference heterogeneity, and as a result, gives the nearest CLV MAPE to our proposed framework. However, where the CRM-only model encounters poor fit is for long-term renewal occasions. The use of fixed-effects on renewal occasions represents an averaging across customer choices at those occasions. Whereas our data fusion framework allows for the evolution of individual-level preference across renewal occasions and can better account for divergent behavior between those who exhibit switching and dropout vs. those who remain on the site as long-term subscribers.

3.6.2 Planned Analyses

The empirical analysis of our study remains on ongoing, and there are several further directions we are presently undertaking to ensure the robustness of our results, which we detail below.

Multinomial Probit (MNP) Specification. A consequence of using the multinomial logit (MNL) model is the assumption that users on the website are choosing across the contract options independently of irrelevant alternatives (IIA). However, as suggested in our model-free evidence, consumers may be viewing alternatives between the short-term subscriptions (e.g., 1-Month) vs. long-term (3- and 6- Month), which may potentially violate the IIA assumption and bias our parameter estimates, unless accounted for. The full error-covariance MNP model (Albert and Chib 1993) can alleviate IIA by allowing for correlations between alternatives over-and-above those accounted for by observables. In testing the MNP against the MNL, we are particularly interested in the difference in accuracy

on the holdout sample CLV due to different specifications. In line with our need for scalability, we are presently developing a stochastic variational EM algorithm for the GPP-MNP specification of our data fusion model.

Latent Class Models. Finite mixture models are a popular method in marketing to approximate heterogeneity distribution across customers (Kamakura and Russell 1989). From a Bayesian perspective, mixture models can represent a flexible way to account for such heterogeneity in data fusion at the customer-level. As such, we view incorporating mixture priors as a natural extension to the preference heterogeneity MNL model already applied in our holdout sample goodness-of-fit exercise.

Time-Trend Only Gaussian Process Priors. Furthermore, we plan to investigate the marginal contribution of the proposed ARD kernel for GPP versus the more widely used square-exponential (SE) and periodic kernels for Gaussian process priors. In particular, we intend to follow Dew and Ansari (2018) by utilizing additive time-trend only GPP over $\{\alpha_{ijt}\}$ in a comparison Bayesian nonparametric fusion framework.

3.7 CONCLUSION

A/B tests and other field experimental techniques have become ubiquitous tools for digital marketers to measure customer responsiveness to, and inform decisions regarding, marketing mix variables, as well as on the design of product and service features. This ubiquity arises from A/B testing's advantage in exposing customers to exogeneously-controlled variations in customers' "natural" process of engagement with the firm. However, in order to measure the full effect of such experimentation, firms often must let tests run for extended periods of time in order to assess the complete temporal trajectory in outcomes-of-interest. This can hamper both timely managerial decision-making and the strategic value of field experimentation for variables, like pricing policy, that have critical long-run downstream consequences.

In this study, we investigate a common A/B test scenario that online subscription services undertake for such a variable: a one-shot pricing experiment on first-time users. Although the experiment is informative by its own merits on the influence of initial pricing on the 'one-shot' upgrade decisions, the revenue

stream for subscription services is critically reliant on resubscriptions. As such, more impactful measurements-of-interest are the effects of various initial pricing regimes on long-run renewal patterns, and by extension, on customer lifetime value. To this end, we develop a Bayesian nonparametric data fusion framework that enables firms to draw on the full data histories of existing customers to help populate the space of potential long-run trajectories of (new) users in the price experiment.

Firms have long availed of other approaches to "fusing" data sets of the type we work with here in order to project test participants' downstream CLV based on information available in the CRM database. Among the most popular of these are: (1) 'naively' matching participants to average CLV values of database customers who've made similar initial purchase choices in the past, (2) matching customers based on propensity scores and nearest neighbor algorithms using observables, or (3) model-based predictions calibrated on existing repeated-choice data in the CRM database. Our framework builds on these approaches by taking an integrated-model approach to fuse the cross-sectional variation of the one-shot experiment to the longitudinal variation of the CRM database to estimate the dynamic evolution of individual-level choice propensity across contract options, $\{\alpha_{ijt}\}$. Such inference is made possible by utilizing a powerful and flexible Bayesian nonparametric prior, the Gaussian process, as a mechanism to differentially share information (e.g., shrinkage) over the space of customer choice occasions, attributes, and time-scales. In comparison to typical matching algorithms map database observations onto experimental ones that serve as input for model-based inference, the data fusion framework jointly integrates optimal mapping of information into model estimation, accounting for statistical uncertainty in both. Moreover, in comparison to models fitted on the database data alone used to predict for the experimental subjects, our framework's fusing of experimental data means that the orthogonalized and wider variation range of the test variable helps to more fully inform downstream temporal inferences (e.g., CLV calculations).

In a holdout sample whose long-run renewal patterns are observed with respect to a set of initial orthogonalized price variations, we find that our proposed framework outperforms matching methods and model-based predictions on either dataset alone, in terms of mean absolute percentage error (MAPE), on customer-level CLV. In particular, for firms seeking to expedite inference on the long-run

effect of A/B testing using data-driven solutions, the proposed method represents a considerable improvement even over a choice model that accounts for preference heterogeneity and time-trends.

Among Bayesian data fusion methods, our framework alleviates two critical bottlenecks of prior work: scalability and the conditional independence assumption (CIA). While our approach can be seen as a generalization of hierarchical Bayes (HB) data fusion by incorporating time dynamics, we additionally now allow for Bayesian shrinkage-based data fusion to be applied to datasets orders of magnitude greater than existing HB approaches allow, and on par with modern database marketing problems. This is made possible by a novel scalable inference strategy based on stochastic variational inference and sparse Gaussian process estimation. Moreover, fundamental to any data fusion method is the CIA assumption in order to form the joint distribution over between datasets. In using the Gaussian process prior as our fusion mechanism, we relax the CIA assumption by avoiding a parametric distribution on the sharing of information across datasets, and instead to allow data points to individually and differentially inform one another. Extending upon both nonparametric and shrinkage-based Bayesian data fusion techniques for discrete-choice models, we have demonstrated that our framework can scalably and accurately applied to a wide range of marketing response and database marketing contexts.

Although the development in the current chapter is squarely focused on the particular setting of fusing one-shot experiments and longitudinal databases, the method underlying this fusion applies far more widely in empirical marketing. The general idea is that of leveraging two or more data sources that each lack a critical feature – analyst control over covariates, imperfect random assignment or representativeness, restricted temporal range, missing covariates or values thereof, insufficient signal variance, covariate multicollinearity, high temporal intercorrelation, etc. – and mitigating those weaknesses by availing of a corresponding strength in the model-informed convex hull of the full data corpus.

A notable feature of this class of methods – in contrast to prior applications of GP in marketing – is the ability to fuse across multiple dimensions at once. As such, the suite of methods can apply with only minor contextual modifications in areas of marketing where customer databases, particularly those containing longitudinal histories, are available. One such area that has received extensive

attention is product recommendations: firms like Amazon have vast histories of the actions and reactions of existing customers, and attempt to leverage sparse information on new customers to make inferences regarding what they might value or purchase. The ability to fuse the results of experiments on new customers to the vast store of actions of existing ones, and to do so across multiple dimensions, should allow results of short-term experiments to extend not only into that customer's future, but to new products, features, or even site architecture. Similarly, retailers who reorganize their physical layout can fuse short-term reactions in market baskets, category incidence, and price tiers to data from their own prior customers and those of other stores in the chain. Such approaches have been applied in specific applications via a nonparametric smoothing approach (e.g., Tank, Foti, and Fox 2015) and can be extended via the sparse variational GP approach to the highly multivariate and multi-data-source settings typical of modern marketing applications.

3.8 TABLES AND FIGURES

Figure 3.1 *Distribution of Time to Initial Conversion*

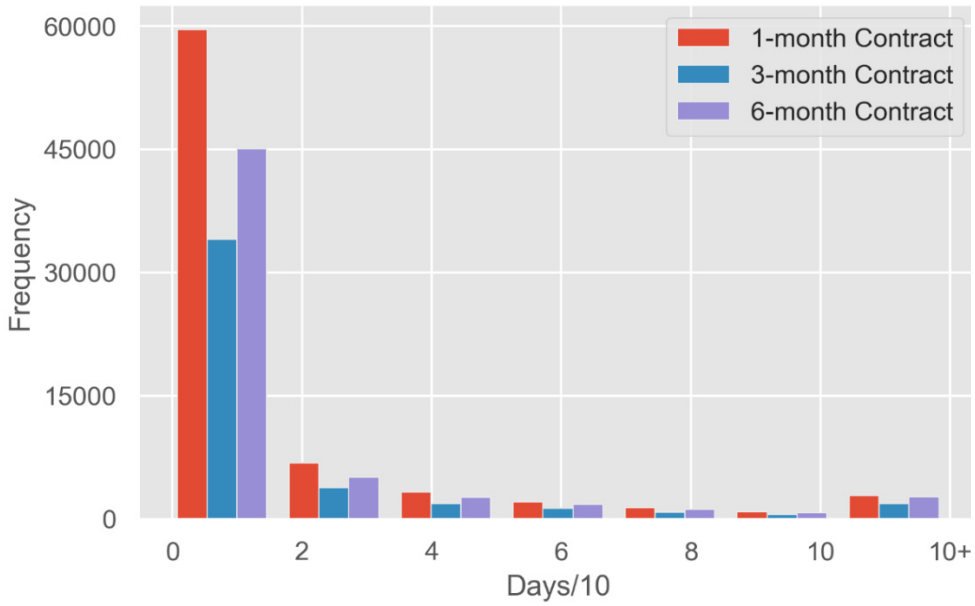


Figure 3.2 *Initial Purchase Incidences by Calendar Weeks*

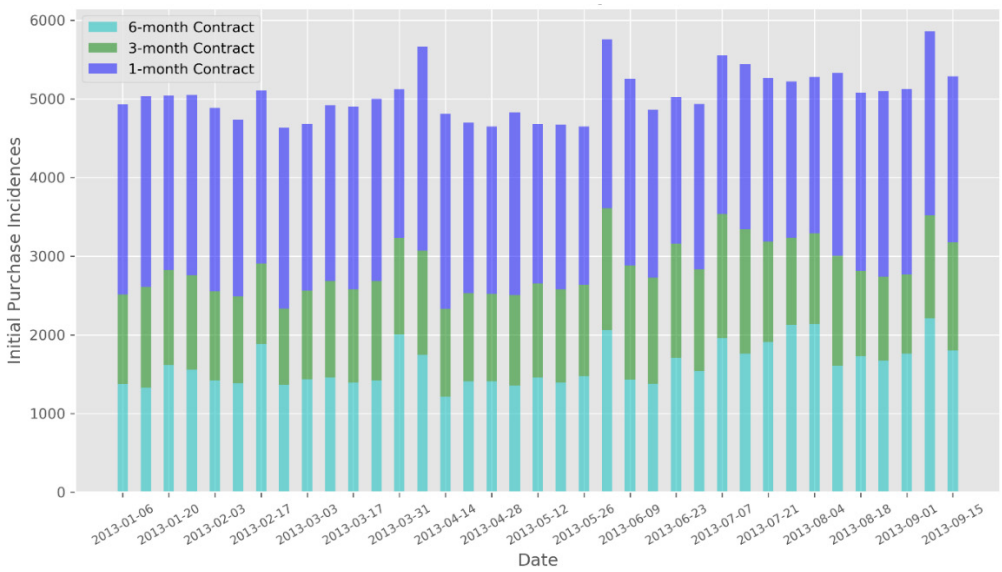


Figure 3.3 *Initial Subscription Choice of First-Time Users*

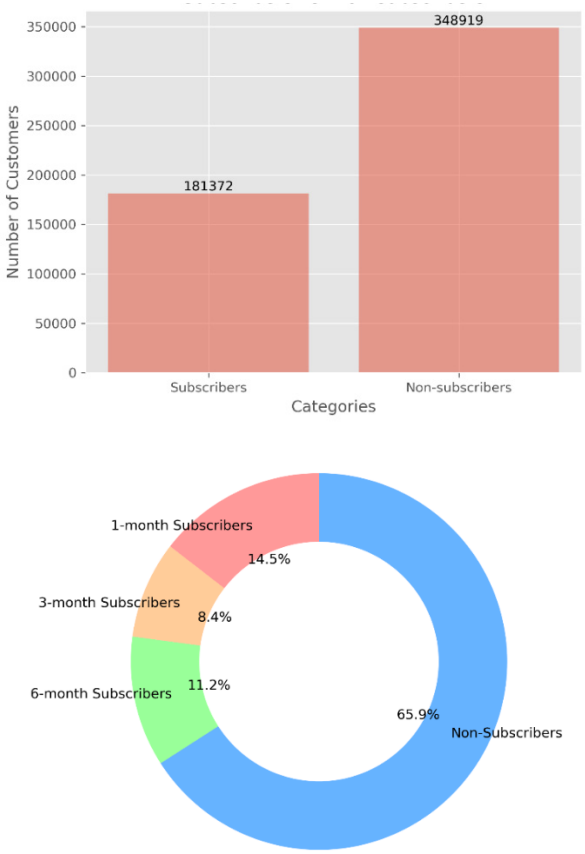


Figure 3.4 *Frequency of Subscription Renewal Occurrences, Users Since Jan. 1*

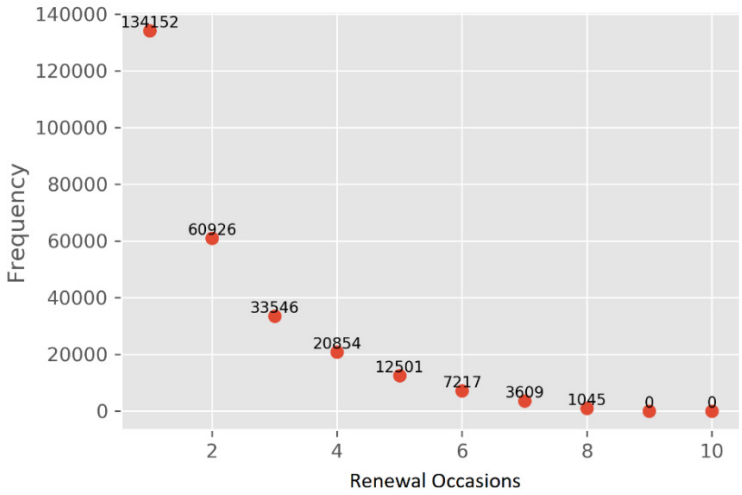
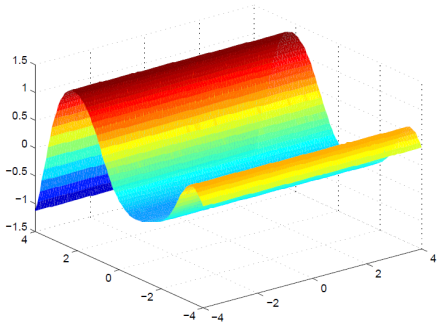
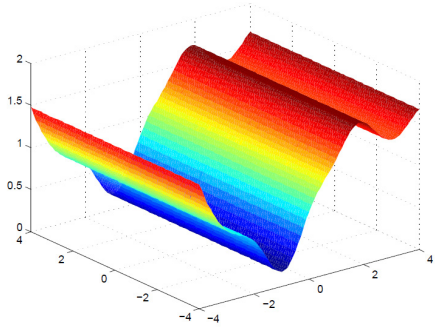


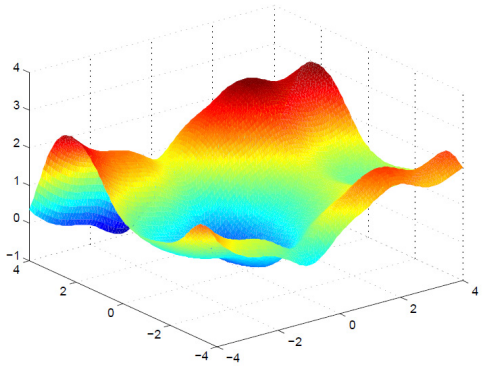
Figure 3.5 *Simulated Gaussian processes*



$k_{SE}(z_1, z_1')$



$k_{SE}(z_2, z_2')$



$k_{ARD}(z, z') = k_{SE}(z_1, z_1') * k_{SE}(z_2, z_2')$

Table 3.1 *Price Conditions, Holdout Sample*

	1-Month ($29.99 standard)	3-Month ($56.97 standard)	6-Month ($83.94 standard)
"High" Condition	$29.99	$53.97	$66.96
"Low" Condition	$26.99	$48.42	$62.96

Table 3.2 *Goodness-of-Fit, Holdout Sample*

Model	*MAPE*
Propensity Score Matching (PSM)	21.7%
Experiment Only (α_j)	33.1%
CRM Only (α_{ij} w/ time FEs)	17.6%
GPP-MNL Data Fusion (α_{ijt})	12.0%

4 Improving Credit Score Forecasts when Data are Sparse: A Dynamic Hierarchical Gaussian Process Model

4.1 ABSTRACT

Credit scores play a vital role in reducing the risk of lending, insuring, and renting to consumers. Credit-based businesses and institutions typically rely on a portfolio of these scores, each informing a specific measure of creditworthiness, in support of decision processes such as vetting prospective customers and setting attractive risk premiums. Centrally problematic to the availability of credit scores is data missingness, which arises from incomplete or inadmissible credit information, otherwise referred to as *thin files*. These missingness can manifest as unscored gaps in the time-series of a consumer's score histories. These gaps reduce the scores' ability to aid in the targeting of profitable borrowers and identifying of cross-selling opportunities for financial products and services. This paper addresses a prevalent form of unscored gaps, referred to as *sparsely unscoreable*, whereby a consumer's credit score portfolio contains periods that contemporaneously exhibit a mix of observed and missing scores. To address this problem, we develop a Bayesian nonparametric latent factor model to impute credible intervals for gaps in individual score histories within a portfolio of dynamically and contemporaneously interrelated scores. We apply this model to novel data from a leading credit bureau on scores for a segment of the U.S. population from 2011-2015, along with attributes derived from credit file used to generate the scores by the credit bureaus. To tackle the high-dimensionality of both the model and feature spaces, we apply the "Gaussian Integral Trick" and the "Reparameterization Trick" to decorrelate prior distributions over latent variables, enabling scalable and accurate model estimation using stochastic mean-field variational inference.

4.2 INTRODUCTION

Assessing the creditworthiness of consumers is a fundamental task for customer acquisition and revenue management by financial institutions (e.g., banks, lenders, and insurers), as well as firms that provide goods and services *on credit* (e.g., utility companies, cellphone plans, rental property managers). The ability of customers to repay loans or make due on rents and payments is vital to the sustainability and profitability of these credit-granting businesses. It's unsurprising then that creditworthiness is most commonly evaluated from the perspective of credit risk; that is, assessing the likelihood of default or delinquency on payment obligations (Anderson 2007). This definition has led to the empirical quantification of credit risk through the use of statistical and decision modeling techniques, a set of approaches that has largely displaced the once-widespread use of "judgmental evaluation" techniques to determine creditworthiness, e.g., a loan officer's subjective assessment of a potential customer (Sullivan 1981).

The process of modeling creditworthiness (credit risk) of individual consumers is referred to as *credit scoring* (Hand and Jacka 1998). Larger institutions have a long history of leveraging in-house data to formulate proprietary credit risk models, of which the earliest variants date back to the 19th century used as part of railroad-shipping ledgers, and today, internal data remains the primary source by banks to estimate probability of default, loss given default, exposure at default and maturity, etc. (Florez-Lopez 2010). Relatively speaking, the standardization and popularization of *commercially available* credit scores in the U.S. arose much later, and can be traced to the emergence of the Federal Fair Credit Reporting Act (1970) and industry-standard FICO scores (1989). Yet, the impact of commercial credit scores on consumer credit markets in the decades since is difficult to overstate. Today, U.S. lenders and credit-granting institutions purchase over 10 billion FICO scores annually to aid in their screening of prospective customers, along with 30 million consumers who access their own scores. According to the Society of Human Resource Management, 45% of companies with over 2,000 employees purchase credit reports from credit bureaus such as Equifax, Experian, and TransUnion, as part of background checks on job applicants. As a result, the three leading U.S. credit bureaus generated over $10 billion in revenue in 2018. In short, "(commercial) credit scoring has been vital in allowing the phenomenal growth in consumer credit over the last five decades" (Thomas et al. 2002).

Of central importance to businesses and employers is the *fidelity* (i.e., predictive accuracy, longitudinal consistency, cross-sectional robustness) of credit scores in classifying potential customers as 'good credit' (i.e., those expected to meet their obligations on a timely basis) vs. 'bad credit' (i.e., those expected to default or be delinquent). These properties of credit scores have received extensive treatment in the literature, in their ability to aid firm-side decisions on customer acquisition (cf. Lee et al. 2002, Sarlija et al. 2004, Lim and Sohn 2007, Sustersic et al. 2009, Hilscher and Wilson 2016, among many), highlighting the integral role these scores play in today's consumer credit markets. Considered as among the most successful statistical modelling approaches used in business applications (Bailey 2004), the scores' fidelity is inherently dependent upon the discriminant power of its input variables. As such, much research on credit scoring models have focused on variable selection across the breadth of data that are made available in credit reports and loan applications. Of those found to be statistically and materially significant on the scores' performance include: primarily, transactional 'tradeline' data such as existing loan amounts, payment and credit history, bank and credit card accounts, outstanding delinquencies and bankruptcy filings (Chen and Huang 2003, Ong et al. 2005); as well as certain less common 'alternative data sources', e.g., time in employment, salary history, cable television accounts (Andreeva 2006, Bellotti and Crook 2009), and more recently, social network data (Wei et al. 2016) [11].

Underlying these extant findings on the key input variables of credit scoring is the presumption that the variables considered are consistently and uniformly available across consumers and over time to businesses and reporting agencies (i.e., required fields in a loan application or mandatory credit report tradelines). However, this is problematic in terms of both (1) limiting *a priori* the feasible set of input attributes to those without missingness, and by extension, (2) the robustness of the credit score models in discerning creditworthiness, particularly if the presence of *nonignorable missingness* cannot be ruled out. For example, in studying consumer loan origination, Orgler (1971) finds that a scoring model built on the information not part of the loan application packet to be more predictive of default than the model built on data arising from the application itself. In their literature review, Abdou and Pointon (2011) highlight that among the scoring models surveyed, the

[11] The Equal Credit Opportunity Act (1974) effectively excludes personal and household demographics and locality information from credit scores.

number of input attributes considered range from as little as three in the most spartan cases to several dozen in more recent work. However, it's worth noting that in all cases, the number of inputs considered remain dwarfed by the several *hundreds* of possible attributes made available from credit reports and agencies. To this, Abdou and Pointon argue that statistically any credit scoring model is incomplete, "but unless a credit scoring model has every possible variable in it, normally it will [lead firms to] misclassify some people."

In this study, we posit credit scores as consumer-time level predictions rising from probabilistic models on the likelihood of default, delinquency, or both (the exact score benchmarks are discussed in Data Description). While a key managerial objective of this study is to improve the fidelity of credit scores, in contrast to prior studies, this study does not propose an alternative credit scoring model. Instead, we investigate a prevalent dilemma faced by firm-side analysts who deal with commercial credit score data with missing score entries (Y_{it}), known to arise due to partially missing input attributes ($X_{it} \coloneqq \{X_{it}^{obs}, X_{it}^{mis}\}$). The analysts' dilemma here is two-fold. First, unbeknownst to the analyst is the functional form relating the former to the latter, as commercial credit score models are trade secrets of credit bureaus (Center for Public Integrity 2011). As a result, the analyst cannot ascertain which set of missing attributes led to a 'unscored' entry. As such, the analyst can *at best* treat the missing set, which may in actuality differ in composition across customers and time periods, as a single amalgamated nonignorable attribute leading to the unscored entry. Taken together, our goal is to then measure the effect of such nonignorable missingness at the customer-time-period level on the fidelity of credit scores and provide a framework to impute scores where previously unscoreable.

To this, we propose a Bayesian nonparametric framework to model the nonignorable amalgams as individual- and time- specific latent factors, regularized by hierarchical Gaussian process (HGP) priors. Taking the perspective that Gaussian processes are distributions over continuous functions, they represent a highly flexible and parsimonious solution to recover the unknown, and likely nonlinear, functional form of the true credit scoring models. To address the heterogeneous composition of the nonignorable amalgams, they are modeled as latent factors drawn hierarchically from customer-specific Gaussian processes indexed over time (months). That is, our HGP framework allows the nonignorable amalgams to

be presumed individual- and time- specific, and their contributed effect to credit scores to have arisen from a broad class of functional forms.

We undertake our investigation on a portfolio of three interrelated credit scores as well as corresponding customer-level credit report attributes provided by a leading U.S. credit bureau, from three metropolitan areas over the sixty months between 2011-2015. While novel to academic settings, our dataset represents a typical product bundle that credit bureaus make available to large credit-granting institutions for purchase. Lending evidence to the prevalence of missing scores, we find that over a quarter (25.8%) of all consumers exhibit unscored entries within their score history.

In the next section, we provide an overview of the dataset as well as the specific form of score missingness this study seeks to tackle. From there, we develop the modeling framework, provide results, and conclude with future extensions.

4.3 DATA DESCRIPTION

FICO and other related commercially available credit scores in the U.S. are issued through four major consumer credit reporting agencies (Equifax, Experian, PRBC, and TransUnion), otherwise referred to as credit bureaus or credit reporting agencies. Most commonly used scores are typically determined exclusively using information available in credit reports[12], although scores leveraging 'alternative data sources' have been developed by credit bureaus as well as the Fair, Isaac, and Co. (FICO). Intended to measure default or delinquency risk over a forward-looking window (typically 12-24 months), commercial credit scores can be directly converted to probabilities using conversion formulas provided by the reporting agencies. Although most credit scores are 'generic' in the sense that they are intended for use by a wide range of creditors, there also exist industry-specific scores (e.g., auto, bankcards, etc.). As such, it's common practice that firms purchase a portfolio of multiple scores to capture the various dimensions of a consumer's credit risk.

Although a non-disclosure agreement (NDA) prevents us from revealing the specific credit bureau and the commercial names of the scores, our data consist of a portfolio of three credit scores from one of the major credit agencies, measuring respectively default, delinquency, and a broader notion of creditworthiness that

[12] https://www.myfico.com/credit-education/credit-scores

combines both default and delinquency risk, all of which span between 2011 and 2015 with monthly observations (i.e., standard cycle for credit score updates). Along with these credit score histories, we are additionally provided with an extensive set of input attributes derived from credit reports, which we detail below. Among the attributes, forty-six of these do not contain any missing values and as such, are incorporated as observed covariates into our model. In applying our analysis to a portfolio of scores as well as credit report attributes, we mimic the common setup typically faced by a firm-side analyst.

4.3.1 Thin vs. Thick Credit Files

As the emphasis of our study is on addressing missing credit score histories, we now provide an overview of the most prevalent source of unscored gaps for individual consumers – namely, *thin file* – which refers to a lack of sufficient history for credit, account, or tradeline activities within an individual consumer's credit report. In addition to insufficiently long credit histories, thin files may also arise as a result of inadmissible inputs due to incomplete or untimely reporting by credit-granting institutions. According to FICO, 53 million Americans in 2016 are considered to have thin files using common credit reporting standards, consisting most commonly of young adults, immigrants, and divorcees. It's worth noting that different commercial credit scores can have different admissibility rules, and as such, the same credit history may be considered a 'thin file' for some scores, but not for others. By extension, this can lead to unscored gaps in only a subset of scores, which we refer to as a 'sparsely unscoreable' credit score portfolio (Fig. 4.1) and a phenomenon in the data we leverage in our imputation framework.

In contrast, scoreable credit histories are referred to a *thick file*. Based on conventional underwriting standards, these are credit files of consumers with at least three credit accounts, with records in all major credit reporting agencies, and having so-called 'mainstream' credit activities within a past 6-month window. These activities typically consist of records of any payments and balances, applications for new accounts, and activities related to existing credit accounts (e.g., mortgages, car and student loans, credit cards). Within the context of a credit report, these activities are referred to and recorded as *tradelines*, which serve as the basis for the input attributes provided in our data. However, there are two key distinctions between credit report tradelines and the attributes provided in our data: (1) our

data have been anonymized to remove all identifying information related to both consumers as well as creditor firms, and (2) the attributes provided contain additional statistics that do not directly appear in credit reports (e.g., lagged sums of collection amounts, min/max age of open trades, and so forth). Moreover, the attributes provided in our data represent the actual set of inputs considered by the credit bureau's score models, although we are not made aware of which attributes are considered by each of the three scores (nor, as mentioned, their actual functional form).

4.3.2 Credit Score Portfolio

The data consist of monthly credit score histories of a portfolio of three scores and associated input attributes for 21.4 million American consumers between 2011 and 2015. Each of the three scores focus on a separate, but interrelated, forecast of credit risk over a forward-looking horizon of 24 months: *bankruptcy* (based on court filings), *delinquency* (obligations unpaid over 90 days), and a *general* measure of creditworthiness (delinquencies and bankruptcies). Across customers, the latter two scores are strongly correlated (0.87), whereas the bankruptcy score has correlations of 0.66 and 0.64 with each of the remaining scores, respectively (Table 4.1). Given the high degree of correlation between the scores, our imputation framework is specified to account for residual correlation (i.e., remainder correlation after controlling for observables) to provide more informed posterior estimates of missing score entries.

In terms of missing entries, we find that 25.8% of all consumers having 'sparse missingness', defined as anytime a consumer has a mix of scored and unscored entries in their score portfolio for a given month. For consumers within our observation period with less than a full-period of credit history (e.g., less than 60 months of observations), the ratio of consumers who are sparsely unscoreable increases to 51.6%. The prevalent occurrence of sparsity in score histories demonstrate the broad applicability of our proposed framework, to be introduced in the next section.

As this study is developed to address sparsely unscoreable consumers, we omit from our analysis any consumers with systematic missingness across all three scores at any point in their score history. Such missingness typically emerge due to structural changes credit history, including (but not limited to) court orders related to bankruptcy resolution, divorce settlements leading to spouses taking on separate

and potentially new credit identities, new immigrants, and children emerging from dependent status who do yet have any credit history. Such structural missingness in credit scores is undoubtedly a fascinating topic, although one which we leave for future research beyond this study. Moreover, for the empirical analysis undertaken in the remainder of this study at present, we direct our focus to a subset of consumers from three metropolitan regions: Detroit ($N = 1685$), Philadelphia ($N = 1355$), and Atlanta ($N = 1567$).

Figures 4.2-4.4 present average score values and sample size by local zip codes across the three scores, overlaid onto the location map of the three metropolitan areas. Perhaps unsurprisingly, these figures illuminate a high degree of spatial correlation of credit scores and neighborhood socio-economics. We find that the more populous but lower-income neighbors within city limits of Detroit and Philadelphia to exhibit lower average credit scores than more affluent suburban zip codes. This trend is less pronounced for Atlanta, which happens to rank among the least economically segregated U.S. metropolitan areas (Diversity and Disparities Project 2019). We will return to a geospatial analysis for our model-based findings, providing evidence that latent drivers of credit scores also exhibit distinct neighborhood-level differences. Taken together, our choice of these three metropolitan regions are designed to capture consumers across representative areas of the U.S.

4.3.3 Input Attributes

We identified 46 attributes from a possible set of 539 provided by the credit bureau to serve as input covariates for the linear component of our imputation model. The filtering rule resulting in the final attribute set is based on (1) having no missing or invalid values across all observations, and (2) to include only those that do not aggregate into other attributes. The latter, by way of an example, implies that among the selected attributes, BANKRUPTCYFLAG is an indicator for whenever a consumer files for a chapter 7, 11, or 13 bankruptcy. Therefore, flags for the individual bankruptcy chapters are excluded for orthogonality and parsimony. Analogous rules are applied to attributes on total balances (e.g., credit cards) that are further broken down into subcategories.

Among the included attributes, they cover six general tradeline categories: age and timing variables (e.g., MONTHSONFILE, AGENEWESTTRADE), account balances (e.g., TOTALBALANCEOPEN, TOTALPASTDUE), collections (e.g., COUNTTAXLIENS,

UNPAIDCOLLECTIONS), flags (BANKRUPTCYFLAG, FORECLOSUREFLAG), inquiries (BYCONSUMER, BYBUSINESSES), as well as trade types (COUNTOPENTRADES, COUNTTRADESSATISFACTORY). A full list of the final attributes is provided in Table 4.2.

4.4 RESULTS

Our analysis was conducted on a sample of 4,607 consumers from zip codes in the Detroit, Atlanta, and Philadelphia metropolitan areas. These consumers are observed to individually have a full 60-month record of all three credit scores. As such, the focus of the following discussion is to provide an understanding of the relationship of observed (input attributes) and latent factors that drive individual-level dynamics across the score portfolio. We additionally revisit the geospatial analysis from Data Description, where strong correlation between neighborhood socioeconomics and score patterns were observed, and extend the analysis to the latent factors. The section ends with a discussion of further data collection needed to move forward with the imputation component of this line of research.

4.4.1 Observed Attributes

We begin by examining patterns among the coefficients of the input attributes, which enter linearly into the SUR framework. Much of the results here service as a 'reality check' on the expected relationship between credit report tradelines and credit scores. Among the 46 input covariates entering into the SUR, all were strongly significant predictors ($p > 0.01$) for at least one of the scores, suggesting that the input attributes selected for the purpose of our analysis represent a core set of attributes utilized by the credit bureau. Furthermore, it's worth noting that in nearly all cases, attribute coefficients are found to take on the same sign for each of the three scores. For example, TOTALPASTDUE and OPENCOLLECTIONS unsurprisingly negatively impact all scores, as any increase in outstanding balances and debt-collection are expected to drive up the likelihood of delinquencies and bankruptcies, to the detriment of the creditworthiness of any consumer. Whereas TRADESSATISFACTORY and AGEOLDESTTRADE, as variables measuring consistent repayment and history of active credit history drive scores upwards, indicating a lower predicted credit risk.

We additionally find that non-significant attributes are strongly centered at zero, which barring a zero-mean polynomial relationship to credit scores, suggest that these attributes are unlikely to have been admitted into a score's calculation. While none of the selected input attributes were nonsignificant for the entire score portfolio, we do find MONTHSONFILE to be *n.s.* (and centered at 0) for the delinquency and bankruptcy scores. This suggests that the length of credit file only plyns a role in determining general creditworthiness, but specialized scores tend to exclude this in favor of actual credit transactions and outcomes. Elucidating on the specialized emphasis of each score, DEROGATORYTRADES a count of trades related to bankruptcy proceedings is *n.s.* for the delinquency scores, as such events arise beyond the tolerance window for delinquent payments; whereas TAXLIENS and THIRDPARTYCOLLECTION are *n.s.* for the bankruptcy score, indicating these are conditions where consumers remain capable of repayment, however untimely.

Lastly, several attributes are observed having opposing signs across scores. Interestingly, while OPENTRADES positively affects general creditworthiness, perhaps signaling a willingness to engage in more credit-based behavior, yet in contrast, having more of these tradelines appears to negatively impact both the delinquency and bankruptcy scores. Additionally, unique to the delinquency score, having more unpaid 3rd party collections (UNPAIDCOLLECTIONS) and more court judgments in public records (COUNTJUDGMENTS) actually increases the delinquency score (e.g., lowing the predicted risk of delinquency). This may be due to that collection agencies and court actions are actions aimed towards ensuring future repayment. Overall, our findings on the input attributes are in-line with common expectations of drivers and their directionality on credit scores, and indeed, may be seen as unsurprising. Moreover, these findings indicate that attributes without missingness are central to most credit scores, as they are consistently available across all consumers. Next, we explore the role of nonlinear, latent drivers of credit scores.

4.4.2 Hierarchical Gaussian Processes

We capture the effect of nonlinear latent factors (γ_{it}^{m}) on credit scores using individual-level Gaussian process priors. While individual- and month- level realizations of γ_{it}^{m} represent the specific estimated effects of the latent factors, these are numerous and difficult to summarize using common summary statistics given their nonlinearity and specificity. Instead this discussion focuses on the two

hyperparameters of the individual-level GPs, the lengthscale (ρ_i^2) and amplitude (η_i^2), to characterize the heterogeneity of latent factors across customers' score portfolios. In our context, the (squared) lengthscale approximates the number of months for consumer i to oscillate between their typical high and low scores, controlling for individual means and attribute effects. As described in the previous section, for each consumer there is a common lengthscale across all three scores. This specification is intended to measure the average intertemporal trajectory of the latent factors across scores in a scale-free fashion (the latter will be captured by the amplitude η_i, which we discuss below). In other words, based on the idea introduced earlier that the latent factors represent the amalgamation of nonlinear 'unobserved' drivers of credit scores, the individual-level lengthscales capture how this amalgam varies over time. It's posited to be individual- and time- specific, but is common across scores, whereas it will be the amplitude that will elucidate on the effect *size* of the amalgam on each of the three scores.

Taking this perspective, we find that the posterior distribution of individual lengthscales to differ across the three geographic regions (Table 4.3). Figure 4.5 overlays the three regions' posteriors, where Detroit is observed to have both the highest average (10.11) and s.d. (11.14) for ρ_i^2, followed by Atlanta (9.22, $s.d. = 9.36$) and Philadelphia (8.41, $s.d. = 7.09$). Overall, the three regions' posteriors are largely overlapping, each with a high degree of dispersion around the mean. The majority of consumers appear to have lengthscales less than 12 months, suggesting that the unobserved component of credit scores most commonly oscillate between typical high and low values within a year, although there exists a significant portion of consumers whose scores tend to be more stable over time (i.e., long right-tail).

Turning our attention to the amplitude $\eta_{i,s}$, this GP hyperparameter captures the customer-specific contribution of latent factors γ_{it}^m to individual scores. As each of the three portfolio scores have distinct ranges and scales, the amplitude captures the score-specific effect size of these latent factors. Figure 4.6 overlays the posterior distribution of the amplitudes by metropolitan area. Here we find that Detroit has a lower degree of dispersion across individual-level amplitudes, compared to Atlanta and Philadelphia. Taken together with the estimates on lengthscale, the latent drivers of Detroit consumers' credit scores appear to oscillate slower between typical highs and lows while exhibiting lower between-consumer heterogeneity in the oscillating range.

Lastly, the consumer-specific cross-score correlation (Σ_i) captures the simultaneous movement cross scores over time while allowing for deviations from the common lengthscale, as described above. As noted in the Data Description, the three scores are highly correlated (Table 4.1). Moreover, we found that the they share many input attributes, often to similar significance and directionality. Table 4.1 additionally shows that after controlling for the observed inputs, the latent factors remain sizable (between 0.28 to 0.39, approximately half of the total correlation). We take this evidence that not only do scores share common observed attributes, but also the so-called 'amalgamated unobservables'. The latter accounts for nearly as much of the correlation across the portfolio as the observables, and as such, lend credence to the need to model these explicitly.

4.4.3 Geospatial Analysis

To better understand the within-region patterns between the focal metropolitan areas, we reconduct the zip code-level analysis, now over the customer-level latent factors. In Detroit (Fig. 4.7), latent factors are observed to exhibit longer lengthscales and lower amplitude among zip codes with high average credit scores. Controlling for the observed covariates, this translates into that these Detroit-area zip codes tend to have high scores and stay high (i.e., low oscillating time-trends), where the reverse is observed for low-scoring zip codes such as those in low-income neighbors within city limits. Similar and more pronounced trends are found for Philadelphia (Fig. 4.8) where affluent and high-scored areas such as downtown Philadelphia and suburbs such as King of Prussia and Berwyn have longer lengthscale and lower amplitude compared to lower- income and score zip codes around north and southwest parts of the city. Comparatively, Atlanta exhibits greater geographic admixture of credit scores (Fig. 4.9), although the inverse relationship between lengthscale and amplitude by zip codes are on-par with Detroit, $corr(\eta_{i,s}, \rho_i^2) = -0.24$. Our findings here are aligned with findings on Atlanta and Detroit being ranked 38th and 33rd in the U.S. terms of most economically segregated commuting regions, whereas Philadelphia is 7th.

This analysis represents a promising approach to studying the intertemporal dynamics of credit scores across communities. The inference on individual-level lengthscales and amplitudes allows a concise characterization on how the evolution

of credit scores differ across cities, zip codes, and of course, individual consumers. Moreover, our results establish our framework's ability to flexibly and robust capture unobservable components of credit scoring, as well as its interplay against observables in driving credit score ratings. Ultimately, our goal is to extend this framework to recover the amalgamated effect of unobservables in sparsely unscored cases, which we now discuss in anticipation of additional data collection.

4.4.4 Missing Data Imputation

As we posit that credit scores are forward-looking predictions arising from probabilistic models taking as inputs the attributes largely corresponding to those we've explicitly selected, in addition to an unknown set of attributes from those we've excluded. Rather than directly imputing individual occurrences of missingness across all possible attributes, almost surely resulting in being beleaguered by the curse of dimensionality, we devised a HGP framework to flexibly impute the amalgamated effect of all excluded attributes, missing or otherwise. The econometrics for imputation is akin to the "omitted variable" perspective of Heckman's (1979) correction, albeit here we avail of a nonparametric prior to share information across scores, individuals, and time periods.

To proceed with the implementation of our 'sparsely unscored' imputation framework in this line of research, we require two additional datasets from our partnering credit bureau. First, we would access score and attribute data on consumers who fit the bill of having 'sparsely unscoreable' portfolios in the 2011-2015 timeframe, augmented with additional 'full-history' consumers akin to the 4,607 considered in this current study. Next, to validate the fidelity of the imputed credit scores, we would access raw tradeline data to uncover actual occurrences of negative credit events (e.g., delinquencies and bankruptcies), and assess the tolerance of the imputed scores to predict these occurrences in (relative) future periods. Should the imputation framework succeed in narrowing the tolerance range (i.e., fidelity) of capturing these events, this would additionally imply that we may augment known score entries without missingness with added information from the imputation mechanism as well.

4.5 CONCLUSION

Unscoreable entries due to thin, incomplete, or inadmissible credit file data are a prevalent phenomenon among commercially available credit scores, and one that commonly bedevils firm-side analysts who are charged with assessing the credit history and creditworthiness of consumers. Rather than proposing a novel credit scoring model – a time-consuming endeavor that risks rendering moot or reinventing existing products and services provided by credit agencies – in this study, we consider the scenario where the analyst may simply wish to impute unscored entries *a posteriori* using existing information from related scores and customers. To this end, we proposed a Bayesian nonparametric framework to provide shrinkage across the information space of customers and scores. Combined with the perspective that credit scores are predictive metrics arising from probabilistic models of unknown functional forms (i.e., a "known unknown"), we posit that if we can flexibly recover the composite effect of the missing inputs, then the missing score must be calculable – i.e., missing scores are not in and of themselves an empirical phenomenon, but rather a consequence. Building on this perspective, we introduced hierarchical Gaussian process priors, as individual-specific *distributions over functions*, to model the composite effect of missing attribute values as arising from a set of customer- and time- specific latent variables. These are made estimable by the HGP's mechanism of 'borrowing information' via shrinkage towards proximate observations. The use of HGP to model latent factors of missingness represents a nonparametric generalization of the Heckman (1979) correction framework for omitted variable bias, and is closely related to work on matrix factorization for missingness in outcomes.

Our framework was applied to a sample of customers from three major U.S. metropolitan areas. At present, the focus of the analysis is to infer the capacity of these latent factors to recover the amalgamated effect of the unobserved drivers of credit scoring, across a portfolio of interrelated scores. We showed that our framework has the ability to capture a wide range of realizations and functional relationships among the latent factors, as demonstrated by the posterior distributions over the HGP lengthscale and amplitude hyperparameters. This enabled a decomposition of credit scores between the effects arising from observables vs. those from unobservables, the latter of which we found to contribute to nearly half of the correlation across a portfolio of score histories, indicative of the

importance to explicitly model the unknowns. Moreover, we undertook a geospatial analysis of credit scores clustered by zip codes, shedding light on how latent factors among customers can inform one another's credit scores.

Taken together, these findings demonstrate the robustness of our framework to be applied to data where missingness are realized and observed. We provided discussion in the previous section on further data collection necessary to extend this line of research. More broadly, the findings here show that credit scoring is a fertile domain for research into consumer finance and response model optimization, and Bayesian nonparametrics as a powerful tool for addressing missing data problems.

4.6 TABLES AND FIGURES

Figure 4.1 *Sparsely unscored gaps, example*

state	zip	archive	m01174	m05143	m05146
FL	32208	201101	0	508	335
FL	32208	201102	0	508	335
FL	32208	201103	0	508	335
FL	32208	201104	0	508	335
FL	32208	201105	0	508	335
FL	32208	201106	0	508	335
FL	32208	201107	0	508	335
FL	32208	201108	0	508	335
FL	32208	201109	0	508	335
FL	32208	201110	0	508	335
FL	32208	201111	0	508	335
FL	32208	201112	0	512	335
FL	32208	201201	0	512	335
FL	32208	201202	0	512	335
FL	32208	201203	0	512	336
FL	32208	201204	0	512	336
FL	32208	201205	0	512	336
FL	32208	201206	0	512	336
FL	32208	201207	0	512	337
FL	32208	201208	0	512	337
FL	32208	201209	0	512	337
FL	32208	201210	0	512	337
FL	32208	201211	0	512	337
FL	32208	201212	594	505	343
FL	32208	201301	594	507	344

Figure 4.2 Scores by Zip Codes, Detroit

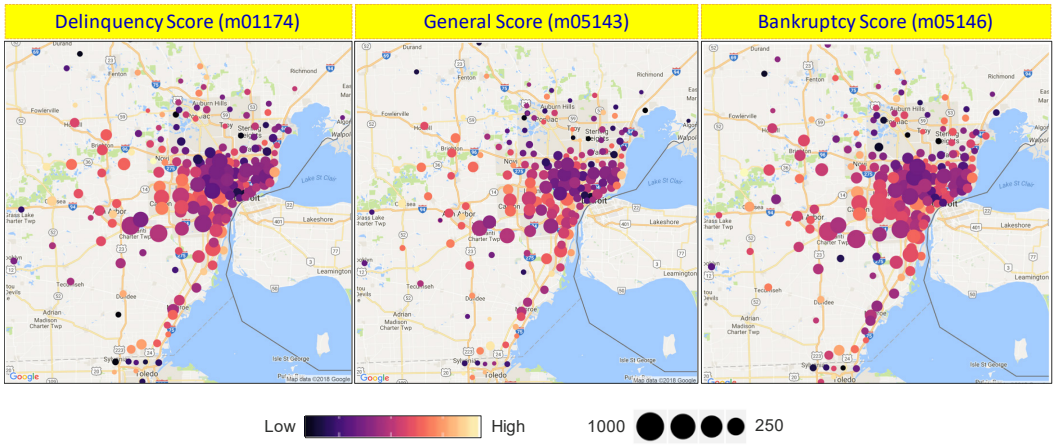


Figure 4.3 Scores by Zip Codes, Atlanta

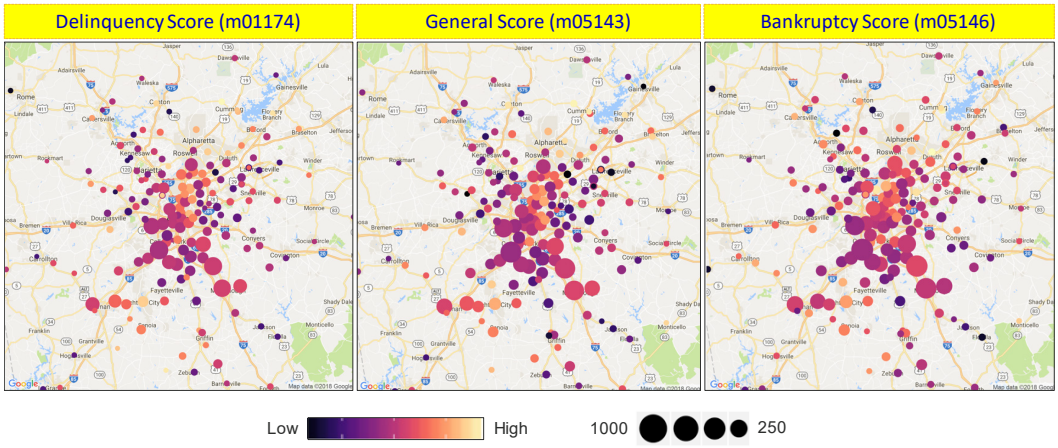


Figure 4.4 Scores by Zip Codes, Philadelphia

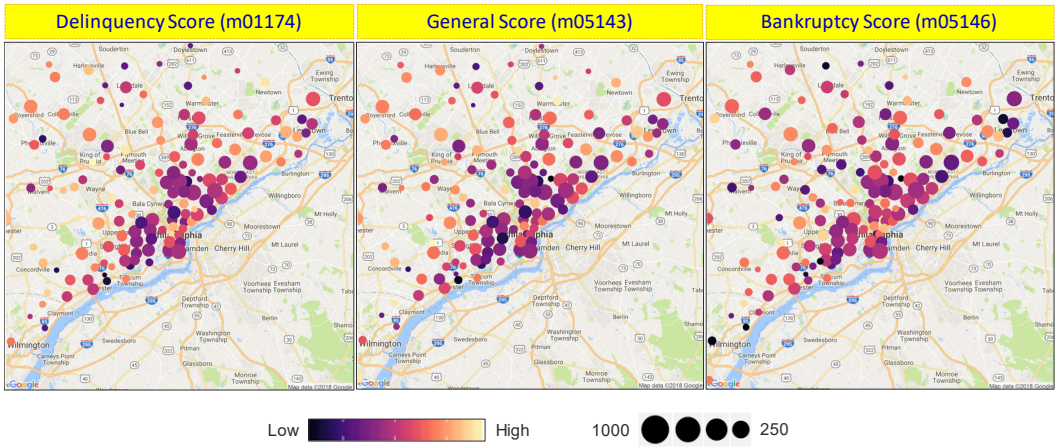


Figure 4.5 *GP Time-trends, Lengthscale* (ρ_i^2)

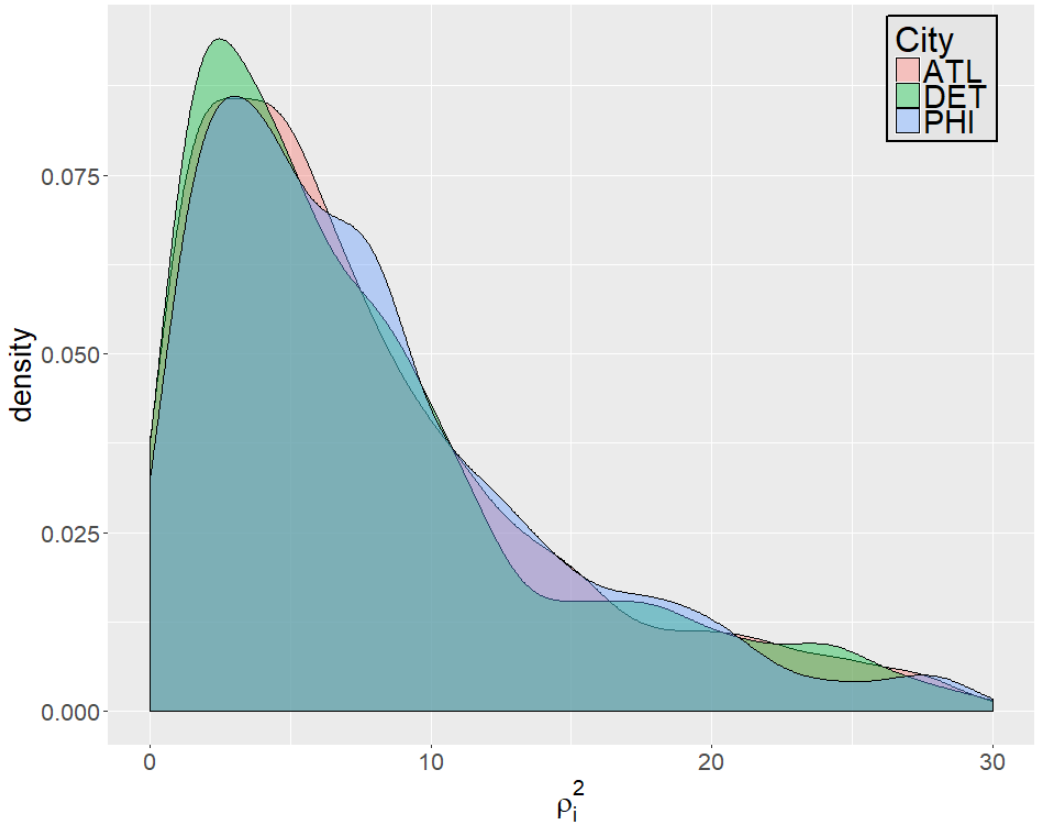


Figure 4.6 *GP Time-trends, Amplitude* (η_i)

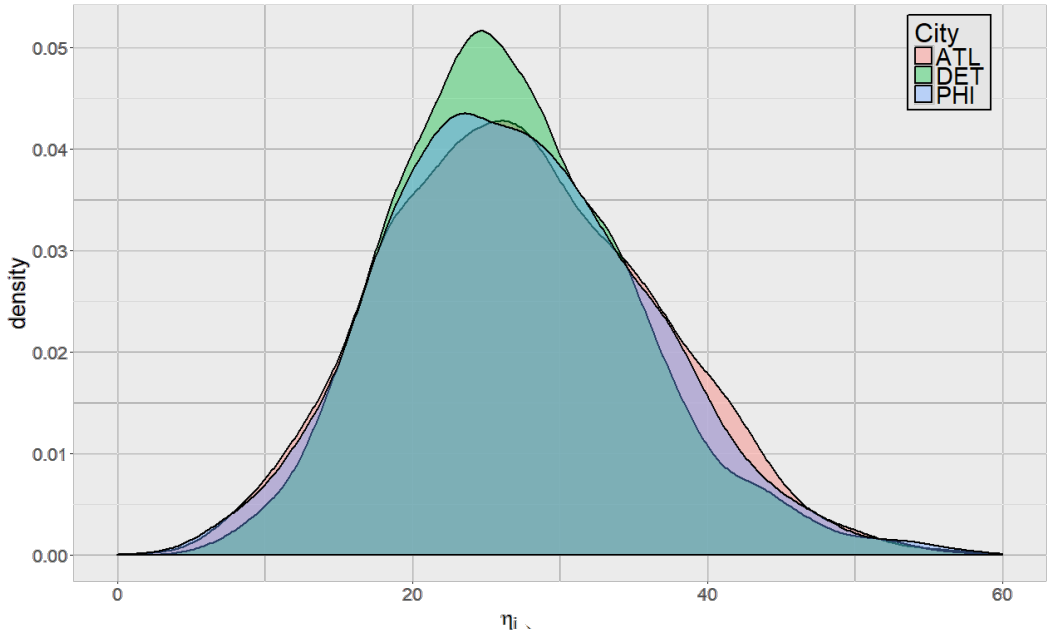


Figure 4.7 GP Time-trends, Detroit

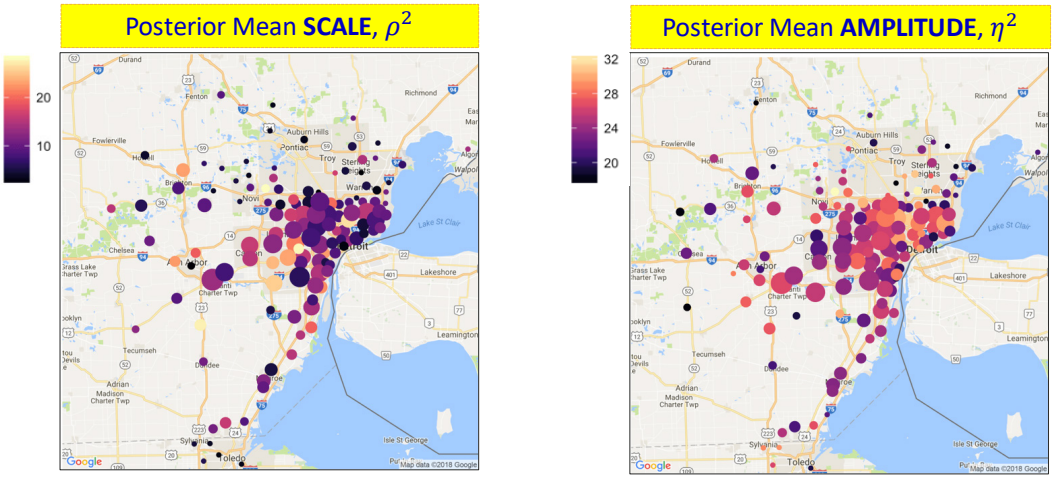


Figure 4.8 GP Time-trends, Philadelphia

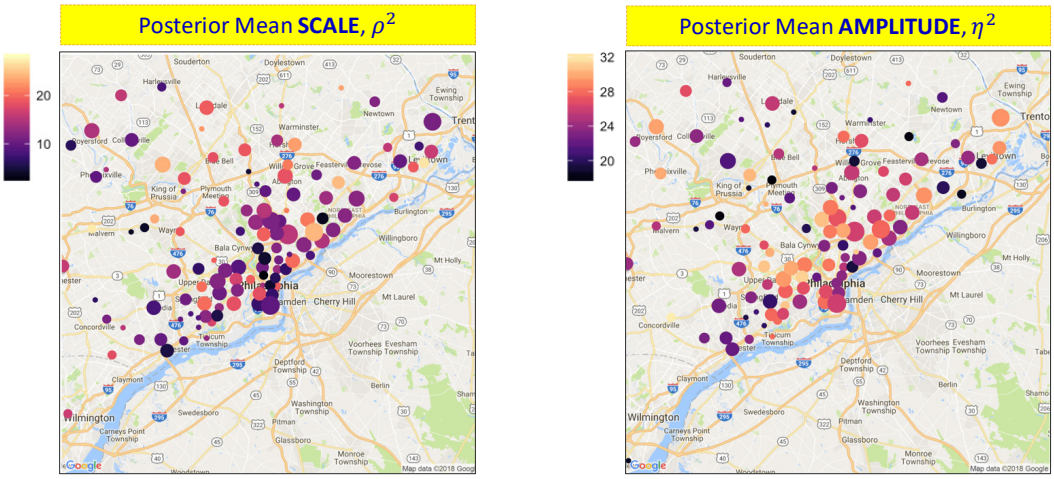


Figure 4.9 GP Time-trends, Atlanta

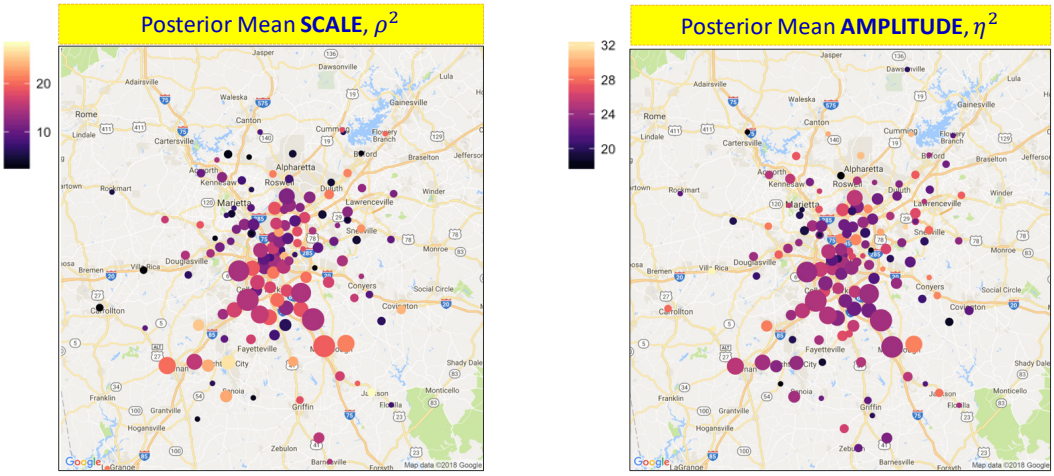


Table 4.1 *Score correlations, full vs. latent factor (latter in parentheses)*

	Delinquency	General	Bankruptcy
Delinquency	1		
General	0.87 (0.39)	1	
Bankruptcy	0.66 (0.34)	0.64 (0.28)	1

Table 4.2 *Input Attribute Descriptions*

Type	Name	Description
Age/Time	ADA_3361	# 60-180 or More Days Past Due Occurrences w/in 12 Months
Age/Time	ADA_3755	Age Newest Date Last Activity Trades Paid as Agreed
Age/Time	ADA_3813	Age Newest Judgment Public Record Item
Age/Time	ADA_3812	Age Newest Tax Lien Public Record Item
Age/Time	ADA_3122	Age Newest Trade
Age/Time	ADA_3111	Age Oldest Trade
Age/Time	ADA_3743	Months on File
Age/Time	ADA_3746	Subject's Age
Balance	ADA_3181	Total Balance Closed Trades w/Update w/in 3 Months
Balance	ADA_3159	Total Balance Open Trades w/Update w/in 3 Months
Balance	ADA_3913	Total Collection Amount 3rd Party Collections
Balance	ADA_3799	Total Collection Amount Unpaid 3rd Party Collections
Balance	ADA_3237	Total Past Due Amount
Collection	ADA_3909	# 3rd Party Collections
Collection	ADA_3815	# of Judgment Pub Rec Item
Collection	ADA_3814	# of Tax Lien Pub Rec Item
Collection	ADA_3807	# Tax Liens, Suits and Judgments, and 3rd Party Collection
Collection	ADA_3796	# Unpaid 3rd Party Collections
Flag	ADA_3903	Bankruptcy Flag
Flag	ADA_3803	Discharged Bankruptcy Public Record Flag
Flag	ADA_3801	Dismissed Bankruptcy Public Record Flag
Flag	ADA_3905	Foreclosure Flag
Flag	ADA_3805	Non-Dismissed, Non-Discharged Bankruptcy Public Record Flag
Flag	ADA_3480	Worst status rating last reported
Inquiries	ADA_3001	# Inquiries w/in 12 Months
Inquiries	ADA_3010	# Non-Utility Inquiries w/in 12 Months
Inquiries	ADA_3030	# Utility Inquiries w/in 12 Months
Trades	ADA_3478	# of trades with status 90-180 DPD last reported
Trades	ADA_3472	# of trades with status satisfactory last reported
Trades	ADA_3137	# Open Trades

Trades	ADA_3100	# Trades
Trades	ADA_3368	# Trades Always Satisfactory
Trades	ADA_3603	# Trades Major Derogatory
Trades	ADA_3135	# Trades Opened w/in 12 Months
Trades	ADA_3338	# Trades Satisfactory w/in 6 Months
Trades	ADA_3625	# Trades Unpaid Major Derogatory
Trades	ADA_3592	# Trades w/ Major Derogatory Event w/in 24 Months
Trades	ADA_3215	# Trades w/ Past Due Amount > $0
Trades	ADA_3614	# Trades w/ Unpaid Major Derogatory Event w/in 24 Months
Trades	ADA_3872	# Trades Worst Rating 120-180+ Past Due or Worse w/in 3 Months
Trades	ADA_3919	# Trades Worst Rating 120-180+ Past Due or Worse w/in 6 Months
Trades	ADA_3448	# Trades Worst Rating 120-180 or More Days Past Due w/in 6 Months
Trades	ADA_3379	# Trades Worst Rating 30 Days Past Due w/in 3 Months
Trades	ADA_3568	# Trades Worst Rating Ever 120-180 or More Days Past Due
Trades	ADA_3952	# Trades Worst Rating Ever 120-180 or More Days Past Due or Worse
Trades	ADA_3888	# Trades Worst Rating No Worse Than 59 Days Past Due w/in 3 Months

Table 4.3 GP Time-trends, Lengthscale (ρ_i^2)

City	Mean	S.D.	2.5%	97.5%
Detroit	10.11	11.14	0.83	43.08
Philadelphia	8.41	7.09	0.24	28.17
Atlanta	9.22	9.36	0.78	36.09

Appendices

APPENDIX A (Essay One)

A.1 – Likelihood Function

For each individual participant in the field experiment, i, we observe a one-shot binary upgrade decision of whether to subscribe during the observation period, and which of the J contracts to choose conditional on upgrade (Eq. 2.2). The individual-level likelihood function is then:

$$l_i(\beta_s, \beta_o, \Sigma) =$$

$$= \sum_{j=0}^{J-1} \underbrace{\Pr\{Y_{s,i} = 1 \cap Y_{o,i} = j\} \boldsymbol{I}\{Y_{s,i} = 1 \cap Y_{o,i} = j\}}_{Upgrade,\ observed\ outcome}$$
$$+ \underbrace{\Pr\{Y_{s,i} = 0 \cap Y_{o,i} = j\} \boldsymbol{I}\{Y_{s,i} = 0 \cap Y_{o,i} = j\}}_{No\ upgrade,\ censored\ outcome}$$

$$= \prod_{j=0}^{J-1} \left[\Pr\{E_{s,i} \cap E_{o,i,j}\}\right]^{d_{s,i} * d_{o,i,j}} \cdot \left[\Pr\{-E_{s,i}\}\right]^{1-d_{s,i}}$$

$$= \prod_{j=0}^{J-1} \left[\int_{E_{s,i} \cap E_{o,i,j}} f_{\varepsilon_i} d\varepsilon_i\right]^{d_{s,i} * d_{o,i,j}} \cdot \left[\int_{-E_{s,i}} f_{\widetilde{\varepsilon}_i} d\widetilde{\varepsilon}_i\right]^{1-d_{s,i}} \quad [A1]$$

Where,

$$d_{s,i} = \boldsymbol{I}\{Y_{s,i} = 1\}$$

$$d_{o,i,j} = \boldsymbol{I}\{Y_{o,i} = j\}$$

$$E_{s,i} = \varepsilon_{s,i} > -X_{s,i}\beta_s$$

$$E_{o,i,j} = \left[\left(\varepsilon_{o,i,j} > -X_{o,i,j}\boldsymbol{\beta_o}\right) \cap \left(\varepsilon_{o,i,j} - \varepsilon_{o,i,k} > \left(X_{o,i,k} - X_{o,i,j}\right)\boldsymbol{\beta_o}\right), \forall k \neq j\right]$$

$$\varepsilon_i = \left(\varepsilon_{s,i}, \varepsilon_{o,i,1}, \dots, \varepsilon_{o,i,J-1}\right), \text{ if } Y_{s,i} = 1$$

$$\tilde{\varepsilon_i} = \left(\varepsilon_{s,i}, \tilde{\varepsilon}_{o,i,1}, \dots, \tilde{\varepsilon}_{o,i,J-1}\right), \text{ if } Y_{s,i} = 0$$

$$f_\varepsilon = \frac{1}{(2\pi)^{D/2}\sqrt{|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}\varepsilon'\boldsymbol{\Sigma^{-1}}\varepsilon\right), \text{ p.d.f. of multivariate normal}$$

When upgrade does not occur ($Y_{s,i} = 0$), the contract choice $Y_{o,i}|Y_{s,i} = 0$ is unobserved – but crucially – assumed to be a censored outcome governed by a latent variable $Y_{o,i}^{*,mis}$ that is correlated with $Y_{s,i}^{*}$, through the error covariance matrix, Σ (Eq. 2.3). As a result, we can integrate over $f_{\tilde{\varepsilon_i}}$, where $\{\tilde{\varepsilon}_{o,i}\}$ are treated as missing conditionally at random (MAR), with respect to $X_{o,i}$, β_o, and Σ, over the domain of $-E_{s,i}$. In comparison, when upgrade is observed, we integrate over f_{ε_i} and over the domain $E_{s,i} \cap E_{o,i,j}$. This enables us to impute these missing values during estimation via data augmentation (Tanner and Wong 1987). By augmenting the error terms of the censored outcomes, this amounts to generating an analogous dataset where missing data are considered as unknown parameters of the model. Note that while we present the estimation procedures below specifically for our candidate model, the technique in applying data augmentation over censored outcomes can be extended with few alterations to all Heckman-type models, treating the estimation of selection models as missing data problems (Zanutto & Bradlow 2006). Taken together, the full sample log-likelihood function is:

$$\mathcal{L}(\boldsymbol{\beta_s}, \boldsymbol{\beta_o}, \boldsymbol{\Sigma}) = \sum_{i=1}^{N}\sum_{j=0}^{J-1} d_{s,i}d_{o,i,j} \ln\left[\int_{E_{s,i}\cap E_{o,i,j}} \frac{1}{(2\pi)^{\frac{1+(J-1)}{2}}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}\varepsilon_i'\boldsymbol{\Sigma^{-1}}\varepsilon_i\right) d\varepsilon_i\right] + \left(1 - d_{s,i}\right) \ln\left[\int_{-E_{s,i}} \frac{1}{(2\pi)^{\frac{1+(J-1)}{2}}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}\tilde{\varepsilon_i}'\boldsymbol{\Sigma^{-1}}\tilde{\varepsilon_i}\right) d\tilde{\varepsilon_i}\right] \quad \text{[A2]}$$

A.2 – Hamiltonian Monte Carlo

As no simple analytic solution exist for Eq. B2, we employ full Bayesian inference to estimate our focal parameters, $(\beta_s, \beta_o, \Sigma)$, using Hamiltonian Monte Carlo (HMC). The key challenge that our estimation strategy overcomes, which has thus far thwarted a tractable approach for selectivity correction for multinomial outcomes, is the difficulty in efficiently sampling from the multivariate normal CDFs in the presence of high-dimensional, correlated missing

data. Our strategy involves sampling the equivalent *data-augmented* posterior over the set of latent utilities $\left(\mathbf{Y}_S^*, \mathbf{Y}_O^*, \mathbf{Y}_O^{*,mis}\right)$, for both censored and uncensored outcomes, and thus obviating the need to solve for the closed-forms of the integrals. We now present the joint posterior density specified in terms of the augmented latent utilities terms (analogous, but in lieu of the integral-based full sample likelihood):

$$p\left(\boldsymbol{\beta_s}, \boldsymbol{\beta_o}, \boldsymbol{\Sigma}, \mathbf{Y}_S^*, \mathbf{Y}_O^*, \mathbf{Y}_O^{*,mis} | \mathbf{Y}_S, \mathbf{Y}_o, \mathbf{X}_s, \mathbf{X}_o\right) \propto \left(\prod_{i=1}^{N} l_i\left(\mathbf{Y}_{S,i}^*, \mathbf{Y}_{O,i}^*, \mathbf{Y}_O^{*,mis} | \boldsymbol{\beta_s}, \boldsymbol{\beta_o}, \boldsymbol{\Sigma}\right)\right) p(\boldsymbol{\beta_s}) p(\boldsymbol{\beta_o}) p(\boldsymbol{\Sigma})$$

In order to formulate the Hamiltonian equations for our sampler, we apply two sets of transformations to ensure valid and efficient HMC sampling: (1) reparametrize the model over $\mathbb{R}^D$ such that the family of the posterior distribution resides on a smooth and differentiable statistical manifold (Neal 2011), and (2) decouple the scale and correlation of the error covariance matrix, Σ, to take advantage of the LKJ correlation prior (Lewandowski et al. 2009).

A.3 – Reparameterization

The decomposition of $\boldsymbol{\Sigma}$ is as follows (Barnard et al. 2000):

$$\boldsymbol{\Sigma} = \boldsymbol{\tau \Omega \tau}$$

Where,

$$\boldsymbol{\tau} = diag\left(\sigma_s, \sigma_{o_1}, \ldots, \sigma_{o_{J-1}}\right)$$

$$\boldsymbol{\Omega} = \begin{pmatrix} 1 & \rho_{s,o_1} & \cdots & \rho_{s,o_{J-1}} \\ & 1 & \cdots & \rho_{o_1,o_{J-1}} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix}$$

Parameters that require transformation to be sampled over $\mathbb{R}$ include the elements of τ and the off-diagonals of Ω. As the elements of τ are strictly positive scale parameters, we utilize an exponential transform $\mathbb{R} \rightarrow \mathbb{R}^+$:

$$\{\sigma_*\} = \{\exp(s_*)\}$$

The off-diagonal elements of the Ω correlation matrix, $\{\rho\} \in (-1, +1)$, must form a positive-definite matrix with diagonal elements equal to 1. To sample over

$\mathbb{R}$ for these elements while ensuring these properties, we utilize a two-step transformation based on Lewandowski et al. (2009). First, for a $K \times K$ correlation matrix, define the set of unconstrained parameters $\{a_*\} \in \mathbb{R}^{\binom{K}{2}}$ transformed by the hyperbolic tangent function, which bijectively maps $\mathbb{R} \to (-1, +1)$, forming the upper triangular elements of the matrix:

$$\mathbf{A} = \begin{pmatrix} 0 & \tanh a_{*,1} & \dots & \tanh a_{*,m} \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \tanh a_{*,M} \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Next, **A** can be transformed into the upper Cholesky triangle **B** of the target correlation matrix, with the following elements, where the target correlation matrix is then given by $\mathbf{\Omega} = \mathbf{B}'\mathbf{B}$:

$$B_{i,j} = \begin{cases} 0 & if\ i > j, \\ 1 & if\ 1 = i = j, \\ \prod_{i'=1}^{i-1} \left(1 - z_{i',j}^2\right)^{\frac{1}{2}} & if\ 1 < i = j, \\ z_{i,j} & if\ 1 = i < j, \\ z_{i,j} \prod_{i'=1}^{i-1} \left(1 - z_{i',j}^2\right)^{\frac{1}{2}} & if\ 1 < i < j \end{cases}$$

A.4 – Latent Utilities and Prior Distributions

In lieu of solving for the high-dimensional integrals found in the full sample likelihood function (Eq. A2), we utilized data augmentation over $\left\{Y_{s,i}^*, \left\{Y_{o,i}^*, Y_{o,i}^{*,mis}\right\}\right\}_{i=}^{N}$, which are realizations from truncated MVN draws with boundary conditions defined by observed upgrade decisions and contract choice outcomes, based on Eq. A1, reproduced below:

$$Y_{s,i}^* = X_{s,i}\beta_s + \varepsilon_{s,i}, \text{ where } Y_{s,i} = \boldsymbol{I}\{Y_{s,i}^* \geq 0\}$$

$$Y^*_{o,i,j} = X_{o,i,j}\beta_o + \sigma_j \varepsilon_{o,i,j} \text{ where } (Y_{o,i}|Y_{s,i} = 1) = \max_j\{Y^*_{o,i,j}\}$$

Additionally, for censored outcomes:

$$Y^{*,mis}_{o,i} = X_{o,j}\beta_o + \tilde{\varepsilon}_{o,j} \text{ where } Y_s = 0 \qquad \text{[A3]}$$

MCMC-based draws for Eq. A3 is asymptotically equivalent to integrating over the set of all possible utility values for the censored outcome, conditional on $X_{o,j}$ and β_o. Taken together, truncated samples of the latent utilities can be drawn according to (superscript refers to the output index of a single truncated MVN distribution):

$$\begin{aligned}
&Y^*_{s,i}|\mathbf{Y_i}, \mathbf{X_i}, \beta_s, \beta_o, \Sigma \sim TMVN^{(1)} \in \min\{0\}, && \text{if } Y_{s,i} = 1 \\
&Y^*_{s,i}|\mathbf{Y_i}, \mathbf{X_i}, \beta_s, \beta_o, \Sigma \sim TMVN^{(1)} \in \max\{0\}, && \text{if } Y_{s,i} = 0 \\
&Y^*_{o,i,j}|\mathbf{Y_i}, \mathbf{X_i}, \beta_s, \beta_o, \Sigma \sim TMVN^{(1+j)} \in \overset{+}{\max}\{Y^*_{o,i,-j}, 0\}, && \text{if } Y_{s,i} = 1 \cap Y_{o,i,j} = j \\
&Y^*_{o,i,j}|\mathbf{Y_i}, \mathbf{X_i}, \beta_s, \beta_o, \Sigma \sim TMVN^{(1+j)} \in \overset{-}{\max}\{Y^*_{o,i,-j}, 0\}, && \text{if } Y_{s,i} = 1 \cap Y_{o,i,j} \neq j \\
&Y^{*,mis}_{o,i}|\mathbf{Y_i}, \mathbf{X_i}, \beta_s, \beta_o, \Sigma \sim TMVN^{(1+j)} \in \mathbb{R}, && \text{if } Y_{s,i} = 0 \qquad \text{[A4]}
\end{aligned}$$

Next, we specify the following priors for the focal parameters, assuming that covariates are standardized and mean-centered to aid in convergence and interpretation of parameters:

$$\begin{aligned}
&\beta_s \sim \mathrm{N}(0,1) \\
&\beta_o \sim \mathrm{N}(0,1) \\
&\tau \sim \mathrm{Cauchy}(0,1) \in \mathbb{R}^{1+(J-1)}_{++} \\
&\Omega \sim \mathrm{LKJ}(1)
\end{aligned}$$

Note that the coefficients (β_s, β_o) are given independent unit normal priors as the latent utilities have relative, rather than absolute, scale with respect to Σ. Any prior variance may suffice without loss of generality, and we thereby choose 1 for ease of interpretation. The strictly positive scale components, τ, are given

uninformative independent, half-Cauchy distributions. The LKJ prior for correlation matrices consists of symmetric Beta distributions with support over $(-1, +1)$. The parameter provided, 1, indicates uniform density over all off-diagonal correlation terms.

A.5 – Stan Implementation

The posterior density in terms of the transformed parameterization, along with the corresponding Jacobians of the transforms, as needed to rescale the priors, is:

$$p\left(\boldsymbol{\beta_s}, \boldsymbol{\beta_o}, \boldsymbol{\tau}, \boldsymbol{\Omega}, \mathbf{Y}_S^*, \mathbf{Y}_O^*, \mathbf{Y}_O^{*,mis} | \mathbf{Y}_S, \mathbf{Y}_o, \mathbf{X}_s, \mathbf{X}_o\right) \propto$$
$$\propto \prod_{i=1}^{N} TMVN\left[\left(Y_{s,i}^*, \left\{Y_{o,i}^*, Y_{o,i}^{*,mis}\right\}\right) | \left(X_{s,i}\beta_s, X_{o,i}\beta_o\right), \Sigma\right] \cdot \prod_{d=1}^{D_S} N(\beta_s | 0,1)$$
$$\cdot \prod_{d=1}^{D_o} N(\beta_o | 0,1) \cdot \prod_{d=1}^{D_\tau} \text{Cauchy}^+(\tau | 0,2) \cdot \text{LKJ}(\Omega | 3) \cdot \left|\frac{\partial \tau}{\partial s_*}\right| \cdot \left|\frac{\partial \Omega}{\partial a_*}\right| \cdot \left|\frac{\partial Y_o^*}{\partial s_*}\right|$$

The final term of the posterior density is unique to the Stan implementation for all truncated and rank-ordered parameters that differ across observations, which in our case are the latent utilities for multinomial choices. As such, truncated outcome utilities (Eq. A4) are implemented in Stan as transformed variables of unbounded parameters. See source code below for details (Algorithm 1).

Algorithm 1 – Stan

```
functions {
  vector f_jj0(vector Zi, int j, int K) {
  // Returns all elements of vector Zi excluding the jth
    vector[K-1] Z_jj0;
    int i = 1;
    for(k in 1:K) {
      if(k != j) {
        Z_jj0[i] = Zi[k];
        i += 1;
      }
    }
    return Z_jj0;
  }
```

```
14
15    real f_max(vector V, real r) {
16    // Returns the max element among a vector and scalar
17       real vmax = max(V);
18       return vmax > r ? vmax : r;
19    }
20 }
21
22 data {
23    int<lower=1> N;                  // no. of total observations
24    int<lower=1> No;                 // no. of total outcomes (i.e., uncensored observations)
25    int<lower=1> Js;                 // no. of binary covariates
26    int<lower=1> Joi;                // no. of outcome covariates, vary across individuals
27    int<lower=1> Jok;                // no. of outcome covariates, vary across choices (i.e., prices)
28    int<lower=1> K;                  // no. of outcome categories minus 1
29    int<lower=1> Oidx[No];           // no. selection-outcome index
30
31    int<lower=0,upper=1> Ys[N];      // binary outcomes
32    int<lower=0,upper=K> Yo[No];     // multinomial outcomes
33    row_vector[Js] Xs[N];            // selection covariates
34    row_vector[Joi] Xoi[N];          // outcome covariates, identical across K
35    matrix[K, Jok] Xok[N];           // outcome covariates, varying across K (i.e., prices)
36 }
37
38 transformed data {
39    int<lower=0> N_pos = sum(Ys);
40    int<lower=0> N_neg = N - N_pos;
41 }
42
43 parameters {
44    // coefficients
45    vector[Js] betas;
46    matrix[Joi, K] betaoi;
47    vector[Jok] betak;
48
49    // binary outcome
50    vector<upper=0>[N_neg] Zs_neg;
51    vector<lower=0>[N_pos] Zs_pos;
52
53    // multinomial outcome
54    vector[K] Zo_ub[No];
55    vector[K] Zo_mis[N-No];
56
57    // error covariance
58    cholesky_factor_corr[K+1] L_omega;
59    vector<lower=0>[K] tau_ub;
60 }
61
62 transformed parameters {
63    vector[K+1] tau = append_row(1, tau_ub);
64    matrix[K+1, K+1] L_sig = diag_pre_multiply(tau, L_omega);
65 }
66
67 model {
68    vector[N] Zs;
69    vector[K] Zo[No];
70    vector[K+1] Z[N];
71
72    // initialize Zs, latent utilities binary selection
73    {  int i = 1; int o = 1;
74       for(n in 1:N) {
75          if(Ys[n] == 1) {
76             Zs[n] = Zs_pos[i];
77             i += 1;
78          } else {
79             Zs[n] = Zs_neg[o];
80             o += 1;
```

```
        }
      }
    }

    // initialize Zo, latent utilities multinomial outcomes
    /* need change-of-variable for ranked bounds, see last line, model section */
    for(i in 1:No) {
      if(Yo[i] == 0) {
        /* bound for y=0: max(Zo) <= 0 */
        Zo[i] = 0 - exp(Zo_ub[i]);
      } else {
        /* bound for chosen: Z_chosen = max(Zo) */
        Zo[i,Yo[i]] = exp(Zo_ub[i, Yo[i]]) + f_max( f_jj0(Zo[i], Yo[i], K), 0);

        /* bound for unchosen: Z_unchosen < Z_chosen */
        for(k in 1:K) {
          if(k != Yo[i]) {
            Zo[i,k] = max([ 0, Zo[i, Yo[i]] ]) - exp(Zo_ub[i,k]);
          }
        }
      }
    }

    // initialize Z, joint latent utilities
    {  int j = 1; int k = 1;
      for (i in 1:N) {
        if(i == Oidx[j]) {
          Z[i] = append_row(Zs[i], Zo[j]);
          if(j < No) j += 1;
        } else {
          Z[i] = append_row(Zs[i], Zo_mis[k]);
          k += 1;
        }
      }
    }

    // OUTCOME LIKELIHOOD //
    for (i in 1:N) {
      vector[K+1] XB = append_row(Xs[i]*betas, Xok[i]*betak + (Xoi[i]*betaoi)');
      Z[i] ~ multi_normal_cholesky(XB, L_sig);
    }

    // PRIORS //
    betas ~ normal(0,1);
    betak ~ normal(0,1);
    to_vector(betaoi) ~ normal(0,1);
    tau_ub ~ cauchy(0,1);
    L_omega ~ lkj_corr_cholesky(1);

    // Change-of-var adjust LL //
    /* log(det(Jacobian)) = log(d(exp(Zo_ub))/dZo_ub) = log(exp(Zo_ub)) = Zo_ub */
    target += Zo_ub;
}

generated quantities {
    vector[Js] Betas = betas;
    matrix[Joi, K] Betao = betaoi/tau_ub[1];
    vector[Jok] Betak = betak/tau_ub[1];

    corr_matrix[K+1] Omega = L_omega*L_omega';
    vector<lower=0>[K-1] Tau = tau_ub;
}
```

APPENDIX B (Essay Two)

We take a step-wise density transform (Ormerod and Wand 2010) strategy in deriving the variational bounds for the MNL and MNP specifications of the GPP data fusion framework. In this strategy, variational distributions for sets of related latent variables, defined as non-focal intermediate parameters, are augmented onto the complete data likelihood. Jensen's inequality is then applied after each augmentation to form the next ELBO until all latent variables are accounted for. The optimal functional forms of the approximating distributions $q(^*)$ are derived using the calculus of variations and Lagrange multipliers, for which we use established solutions from existing literature.

B.1 – Polytomous Choice Utilities

In deriving the ELBO of the GPP data fusion MNL and MNP specifications, we note that they share a common random utility framework consisting of a deterministic component ψ_{ijt} and an additive noise ε_{ijt} where the observed choice outcome y_{ijt} represents the choice alternative with the highest total utility,

$$\varepsilon_{ijk} \sim \phi(\cdot),\ j \in \{1, \dots, J\}$$

$$y_{ijt} = \underset{\mathrm{j}}{\operatorname{argmax}}\left(\psi_{ijt} + \varepsilon_{ijt}\right) \qquad \text{[B1]}$$

The MNL specification arises when assuming the error distribution $\phi(\cdot)$ is the Extreme Value Type I (EV-1), whereas the MNP arises assuming a (multivariate) normal. In relation to the generalized data fusion utility structure (Eq. 3.6a) then:

$$\psi_{ijt} = \alpha_{ijt} + p_{ijt}\beta + X_{it}\gamma_j \qquad \text{[B2]}$$

Under the random utility framework, the marginal probability of choosing alternative j is equivalent to the probability that its total utility $\psi_{ijt} + \varepsilon_{ijt}$ being greater than all other alternatives. Let $\Phi(\cdot)$ denote the corresponding error CDF,

$$p(y_{ijt} = j|\psi_{ijt}) = \Pr(\psi_{ijt} + \varepsilon_{ijt} \geq \psi_{ij't} + \varepsilon_{ij't}, \forall j' \neq j) \quad [B3]$$
$$= \int_{-\infty}^{+\infty} \phi(\varepsilon) \prod_{j' \neq j} \Phi(\varepsilon + \psi_{ijt} - \psi_{ij't})\, d\varepsilon \triangleq \boldsymbol{\mathcal{L}_o}$$

In line with subsequent terminology, we refer to the data likelihood as the zeroth or starting bound ($\boldsymbol{\mathcal{L}_o}$). Ruiz et al. (2018) note that from this form of the marginal likelihood, the softmax solution of the MNL and the data augmentation estimation strategy for MNP can be derived.

B.2 – Multinomial Logit

B.2.1 One-vs-Each Bound

The first bound on the MNL specification is based on the one-vs-each bound derived in Titsias (2016). In solving Eq. B3 with EV-1 errors, the multinomial logit CDF has the form of the softmax function,

$$p(y_{ijt} = j|\psi) = \frac{\exp(\psi_{ijt})}{\sum_{j=1}^{J} \exp(\psi_{ij't})} \quad [B4]$$

This can be rewritten as,

$$p(y_{ijt} = j|\psi) = \frac{1}{1 + \sum_{j' \neq j}^{J} \exp\left(-(\psi_{ijt} - \psi_{ij't})\right)} \quad [B5]$$

A lower bound on the softmax probability can be given by a closely identical form known as the one-vs-each softmax, which involves replacing the summation in the denominator of Eq. B5 with a product over sigmoid functions,

$$p(y_{ijt} = j|\psi) \geq \prod_{j' \neq j} \frac{1}{1 + \exp\left(-(\psi_{ijt} - \psi_{ij't})\right)} \qquad [B6]$$
$$= \prod_{j' \neq j} \sigma(\psi_{ijt} - \psi_{ij't}) \triangleq \mathcal{L}_1$$

B.2.2 Pólya-Gamma Data Augmentation

Bayesian inference directly on logistic regressions does not result in known forms on all full conditional distributions of the posterior. This has led to the Metropolis-Hastings algorithm as the most prevalent posterior inference approach for the MNL, in contrast to the most efficient (Gaussian) data augmentation-based Gibbs sampler for the MNP (Albert and Chib 1993, McCulloch and Rossi 1994). Polson et al. (2013) posited the errors of binomial distributions, including the Gumbel and standard logistic, as a scale mixture of Gaussians under the Pólya-Gamma (PG) distribution that admits a data augmentation Gibbs sampler composed of Gaussian draws for linear coefficients and PG draws for a single-layer of latent variables. At the heart of the PG distribution is the equivalence of a generalized sigmoid function to the expectation over the PG random-variate scaled quadratic exponential (e.g., Gaussian) term,

$$\frac{(\exp(\psi))^a}{(1 + \exp(\psi))^b} = 2^{-b} \int_0^\infty \exp\left(\kappa\psi - \frac{\omega\psi^2}{2}\right) p(\omega) d\omega \qquad [B7]$$

For logistic regressions[13], $a, b = 1$, $\kappa = y_j - 1/2$ for $y_j \in \{-1, 1\}$ denoting whether the choice outcome was alternative j, and $\omega \sim PG(1, 0)$. The distributions of ω are given by,

$$PG(\omega|b, 0) \propto \frac{2^{b-1}}{\Gamma(b)} \sum_{n=0}^{\infty} (-1)^n \frac{\Gamma(n+b)}{\Gamma(n+1)} \frac{(2n+b)}{\sqrt{2\pi\omega^3}} \exp\left(-\frac{(2n+b)^2}{8\omega}\right) \quad \text{[B8a]}$$

$$PG(\omega|b, c) = \frac{\exp(-c^2\omega/2) PG(\omega|b, 0)}{\mathbb{E}_\omega[\exp(-c^2\omega/2)]} \quad \text{[B8a]}$$

It's worth noting that although variational Bayes differs from the sampling-based inference strategy of MCMC, the identity from Eq. B7 that enables a data augmentation Gibbs sampler due to conditional conjugacy likewise enables the formulation of an efficient single-bound closed-form solution for variational inference on GPP logistic regressions (Wenzel et al. 2018). Our formulation of the GPP MNL variational bound builds on the binary logit specification of Wenzel et al. (2018) who utilized the PG augmented representation of the sigmoid function (Eq. B7) to admit a single-bound ELBO with analytical natural gradients for efficient posterior inference. To adapt the PG data augmentation for MNL, we apply the strategy of Polson et al. (2013) in reformulating the categorical softmax function in terms of sigmoid functions. Specifically, we utilize the one-vs-each $\boldsymbol{\mathcal{L}_1}$ (Eq. B6) for the softmax to proceed in generalizing Wenzel et al.'s (2018) approach to MNL, which results in:

$$\begin{aligned} \boldsymbol{\mathcal{L}_1} &= \prod_{j' \neq j} \sigma\left(\psi_{ijt} - \psi_{ij't}\right) \\ &= \prod_{j' \neq j} \frac{1}{2} \int_0^{\infty} \exp\left(\frac{\left(\psi_{ijt} - \psi_{ij't}\right)}{2} - \frac{\left(\psi_{ijt} - \psi_{ij't}\right)^2}{2} \omega_{ijt}\right) p(\omega) d\omega \end{aligned} \quad \text{[B9]}$$

[13] See Polson et al. (2013) section 3 for proof.

As noted, the central effort of variational Bayes is to map intractable integrals as tractable optimization through density transforms. To tackle the integrals found in Eq. B9 for which no analytical solution is known, we augment this joint density with auxiliary variational densities for the Pólya-Gamma variates ω. Collecting terms over observations, the augmented joint density w.r.t to $\boldsymbol{\mathcal{L}_1}$ results in the integral-free form of,

$$p(\boldsymbol{y}, \boldsymbol{\psi}, \boldsymbol{\omega}) = p(\boldsymbol{y}|\boldsymbol{\psi}, \boldsymbol{\omega})p(\boldsymbol{\psi})p(\boldsymbol{\omega})$$
$$\propto \prod_{j' \neq j}^{J} \exp\left[\frac{1}{2}\left(\boldsymbol{\psi}_j - \boldsymbol{\psi}_{j'}\right) - \frac{1}{2}\left(\boldsymbol{\psi}_j - \boldsymbol{\psi}_{j'}\right)^{\mathrm{T}} \Omega \left(\boldsymbol{\psi}_j - \boldsymbol{\psi}_{j'}\right)\right] p(\boldsymbol{\psi})p(\boldsymbol{\omega}) \quad \text{[B10]}$$

where Ω is the diagonal matrix of the PG variables $\{\omega_{ijt}\}$. Crucially, in contrast to the original MNL specification, the PG augmented form is amenable to conditional conjugacy over both the Gaussian process and linear terms within ψ that enables the derivation of closed-form updates. Moreover, the term $p(\boldsymbol{y}|\boldsymbol{\psi}, \boldsymbol{\omega})$ can be thought of as the PG variate augmented form of the marginal likelihood $p(y = j|\psi)$ (Eq. B6) with only difference due to the inequality introduced by $\boldsymbol{\mathcal{L}_1}$. Finally, in line with standard identification constraints for polytomous choice models, ψ_j for the baseline alternative (e.g., 'No Subscription') is held to zero.

B.2.3 Sparse GP for MNL

Reproducing the optimal augmented distribution of $\boldsymbol{\alpha}$ and variational distribution of $\mathbf{u}$ given in section 4.2,

$$p(\boldsymbol{\alpha}|\boldsymbol{u}) = MVN\left(\boldsymbol{\alpha} \mid K_{nm}K_{mm}^{-1}\boldsymbol{u}, \widetilde{K}\right) \quad \text{[B11]}$$

$$p(\boldsymbol{u}) = MVN(\boldsymbol{u} \mid 0, K_{mm}) \quad \text{[B12]}$$

Incorporated into the PG-augmented joint density w.r.t. to $\boldsymbol{\mathcal{L}_1}$ (Eq. B10) results in,

$$p(\boldsymbol{y}, \boldsymbol{\psi}, \boldsymbol{\omega}, \boldsymbol{u}) = p(\boldsymbol{y}|\boldsymbol{\psi}, \boldsymbol{\omega})p(\boldsymbol{\omega})p(\boldsymbol{\psi}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}|\boldsymbol{u})p(\boldsymbol{u}) \quad \text{[B13]}$$

B.2.4 Multinomial Logit: Complete Variational Bound

To formulate the final variational bound on the MNL specification of the GPP data fusion framework, we begin by applying the Jensen's inequality on the conditional log-likelihood of the observed choice outcomes,

$$\begin{aligned} \log p(\boldsymbol{y}|\boldsymbol{\psi}, \boldsymbol{\omega}, \boldsymbol{u}) &= \log \mathbb{E}_{p(\boldsymbol{\alpha}|\boldsymbol{u})}[p(\boldsymbol{y}|\boldsymbol{\psi}, \boldsymbol{\omega})] \\ &\geq \mathbb{E}_{p(\boldsymbol{\alpha}|\boldsymbol{u})}[\log p(\boldsymbol{y}|\boldsymbol{\psi}, \boldsymbol{\omega})] \end{aligned} \quad \text{[B14]}$$

Introducing the variational distributions of the inducing points $q(\boldsymbol{u})$ and the augmented PG variates $q(\boldsymbol{\omega})$ on this inequality on the likelihood marginalizing over all parameters results in the ELBO,

$$\begin{aligned} \log p(\boldsymbol{y}) \geq \mathbb{E}_{p(\boldsymbol{a}|\boldsymbol{u})q(\boldsymbol{u})q(\boldsymbol{\omega})}[\log p(\boldsymbol{y}|\boldsymbol{\psi}, \boldsymbol{\omega})] - KL(q(\boldsymbol{u})||p(\boldsymbol{u})) \\ - KL(q(\boldsymbol{\omega})||p(\boldsymbol{\omega})) \end{aligned} \quad \text{[B15]}$$

where $q(\omega_{ijt}) = PG(\omega_{ijt} \mid 1, c_{ijt})$ and $q(\boldsymbol{u}) = MVN(\boldsymbol{u} \mid \boldsymbol{\mu_u}, \Sigma_{\mathrm{u}})$. Note that the variational distributions for $\boldsymbol{\omega}$ and $\boldsymbol{u}$ introduce an additional layer of local (observation-level) parameters $\{c_{ijt}\}$ and global parameters $\boldsymbol{\mu_u}, \Sigma_{\mathrm{u}}$ that now enter into the ELBO. These parameters enable the decomposition of the optimization updates of the ELBO objective function across minibatches data to allow for

scalable stochastic variational inference (Hoffman et al. 2013, Wenzel et al. 2018). Solving for Eq. B14 using the identity from Eq. B11 results in,

$$\begin{aligned}
&\mathbb{E}_{p(\boldsymbol{\alpha}|\boldsymbol{u})}[\log p(\boldsymbol{y}|\boldsymbol{\psi},\boldsymbol{\omega})] \\
&\quad = \mathbb{E}_{p(\boldsymbol{\alpha}|\boldsymbol{u})}\left[\sum_{j'\neq j}\frac{1}{2}(\boldsymbol{\psi}_j - \boldsymbol{\psi}_{j'}) \right. \\
&\quad \left. -\frac{1}{2}(\boldsymbol{\psi}_j - \boldsymbol{\psi}_{j'})^{\mathrm{T}}\Omega(\boldsymbol{\psi}_j - \boldsymbol{\psi}_{j'})\right] \\
&\quad = \frac{1}{2}\sum_{j\neq j'}\left(\Upsilon - \Upsilon^{\mathrm{T}}\Omega\Upsilon - tr\left(\Omega\widetilde{\mathrm{K}}_j + \Omega\widetilde{\mathrm{K}}_{j'}\right)\right)
\end{aligned} \tag{B16}$$

where $\Upsilon = \left(K_{jnm}K_{jmm}^{-1}\boldsymbol{u_j} + X\beta_j\right) - \left(K_{j'nm}K_{j'mm}^{-1}\boldsymbol{u_{j'}} + X\beta_j\right)$ is the difference in mean functions of the inducing point distribution, along with the linear coefficients. The trace terms from Eq. B16 arise from the covariance term of Eq. B14, representing the logged scale of the inducing point distribution. Incorporating this result into Eq. B15 (excluding the KL divergence terms),

$$\begin{aligned}
&\mathbb{E}_{p(\boldsymbol{a}|\boldsymbol{u})q(\boldsymbol{u})q(\boldsymbol{\omega})}[\log p(\boldsymbol{y}|\boldsymbol{\psi},\boldsymbol{\omega})] \\
&\quad = \frac{1}{2}\mathbb{E}_{q(\boldsymbol{u})q(\boldsymbol{\omega})}\left[\sum_{j\neq j'}\left(\Upsilon - \Upsilon^{\mathrm{T}}\Omega\Upsilon \right.\right. \\
&\quad \left.\left. - tr\left(\Omega\widetilde{\mathrm{K}}_j + \Omega\widetilde{\mathrm{K}}_{j'}\right)\right)\right] \\
&\quad = \frac{1}{2}\mathbb{E}_{q(\boldsymbol{u})}\left[\sum_{j\neq j'}\left(\Upsilon - \Upsilon^{\mathrm{T}}\Xi\Upsilon - tr\left(\Xi\widetilde{\mathrm{K}}_j + \Xi\widetilde{\mathrm{K}}_{j'}\right)\right)\right] \\
&= \frac{1}{2}\left[\sum_{j\neq j'}\left(\widetilde{\Upsilon} - \widetilde{\Upsilon}^{\mathrm{T}}\Xi\widetilde{\Upsilon} - tr\left(\Xi\widetilde{\mathrm{K}}_j + \Xi\widetilde{\mathrm{K}}_{j'}\right)\right.\right. \\
&\qquad \left.\left. - tr\left(\boldsymbol{\kappa}_j^{\mathrm{T}}\Xi\boldsymbol{\kappa}_j\Sigma_{uj} + \boldsymbol{\kappa}_{j'}^{\mathrm{T}}\Xi\boldsymbol{\kappa}_{j'}\Sigma_{uj'}\right)\right)\right]
\end{aligned} \tag{B17}$$

where,

$$\Xi = diag(\{\xi_{ijt}\}),\ \xi_{ijt} = \mathbb{E}_{p(\omega_i)}[\omega_{ijt}] = \frac{1}{2c_{ijt}}\tanh\left(\frac{c_{ijt}}{2}\right)$$

$$\widetilde{\Upsilon} = \mathbb{E}_{p(\boldsymbol{u})}[\Upsilon] = \left(K_{jnm}K_{jmm}^{-1}\boldsymbol{\mu_{uj}} + X\beta_j\right) - \left(K_{j\prime nm}K_{j\prime mm}^{-1}\boldsymbol{\mu_{uj\prime}} + X\beta_j\right)$$

$$\boldsymbol{\kappa}_j = K_{jnm}K_{jmm}^{-1}$$

Next we derive the analytical form of the KL divergence for $\boldsymbol{u}$ and $\boldsymbol{\omega}$. As both $q(u)$ and $p(u)$ are multivariate Gaussian, the KL divergence has the known closed-form solution of (Duchi 2007),

$$\begin{aligned} KL\left(q(\boldsymbol{u_j}) \,||\, p(\boldsymbol{u_j})\right) \\ = \frac{1}{2}\left(tr\left(K_{jmm}^{-1}\Sigma_{\mathrm{j}}\right) + \boldsymbol{\mu}_{uj}K_{jmm}^{-1}\right. \\ \left. - \log\left|\Sigma_{uj}\right| + \log|K_{mm}|\right) \end{aligned} \quad [B18]$$

In line with Wenzel et al. (2018), the KL divergence for ω is,

$$\begin{aligned} KL\left(q(\boldsymbol{\omega}) \,||\, p(\boldsymbol{\omega})\right) &= \mathbb{E}_{q(\boldsymbol{\omega})}[\log q(\boldsymbol{\omega}) - \log p(\boldsymbol{\omega})] \\ &= \sum^{I,J,T} \log\cosh\left(\frac{c_{ijt}}{2}\right) - \frac{c_{ijt}}{4}\tanh\left(\frac{c_{ijt}}{2}\right) \end{aligned} \quad [B19]$$

Finally, summing over all terms and solving for Eq. B15 the final ELBO for the MNL specification of the GPP data fusion framework is,

$$\begin{aligned} \boldsymbol{\mathcal{L}}(\boldsymbol{c}, \{\boldsymbol{\mu}_u\}, \{\boldsymbol{\Sigma}_u\}) = \frac{1}{2}\Bigg(& \sum_{j \neq j'}\Big(\widetilde{\Upsilon} - \widetilde{\Upsilon}^{\mathrm{T}}\Xi\widetilde{\Upsilon} - tr\left(\Xi\widetilde{\mathrm{K}}_j + \Xi\widetilde{\mathrm{K}}_{j'}\right) \\ & - tr\left(\boldsymbol{\kappa}_j^{\mathrm{T}}\Xi\boldsymbol{\kappa}_j\Sigma_{uj} + \boldsymbol{\kappa}_{j\prime}^{\mathrm{T}}\Xi\boldsymbol{\kappa}_{j\prime}\Sigma_{uj\prime}\right)\Big) \\ & - \sum^{I,J,T}_{J} 2\log\cosh\left(\frac{c_{ijt}}{2}\right) - \frac{c_{ijt}}{2}\tanh\left(\frac{c_{ijt}}{2}\right) \\ & - \sum_{j=1} tr\left(K_{jmm}^{-1}\Sigma_{\mathrm{j}}\right) + \boldsymbol{\mu}_{uj}K_{jmm}^{-1} \\ & - \log\left|\Sigma_{uj}\right| + \log|K_{mm}|\Bigg) \end{aligned} \quad [B20]$$

Bibliography

Abdou, Hussein A., and John Pointon. "Credit scoring, statistical techniques and evaluation criteria: a review of the literature." *Intelligent Systems in Accounting, Finance and Management* 18.2-3 (2011): 59-88.

Adigüzel, Feray, and Michel Wedel. "Split questionnaire design for massive surveys." *Journal of Marketing Research* 45.5 (2008): 608-617.

Albert, James H., and Siddhartha Chib. "Bayesian analysis of binary and polychotomous response data." *Journal of the American statistical Association* 88, no. 422 (1993): 669-679.

Allenby, Greg M., Thomas S. Shively, Sha Yang, and Mark J. Garratt. "A choice model for packaged goods: Dealing with discrete quantities and quantity discounts." *Marketing Science* 23, no. 1 (2004): 95-108.

Anderson, Eric T., and Duncan I. Simester. "Long-run effects of promotion depth on new versus established customers: Three field studies." *Marketing Science* 23, no. 1 (2004): 4-20.

Anderson, Raymond. *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation.* Oxford University Press, 2007.

Andreeva, Galina. "European generic scoring models using survival analysis." *Journal of the Operational Research Society* 57.10 (2006): 1180-1187.

Andridge, Rebecca R., and Roderick JA Little. "A review of hot deck imputation for survey non-response." *International statistical review* 78.1 (2010): 40-64.

Ansari, Asim, Yang Li, and Jonathan Z. Zhang. "Probabilistic Topic Model for Hybrid Recommender Systems: A Stochastic Variational Bayesian Approach." *Marketing Science* 37.6 (2018): 987-1008.

App Association, "State of the App Economy", 6th Edition (2018).

Ascarza, Eva, Anja Lambrecht, and Naufel Vilcassim. "When talk is "free": The effect of tariff structure on usage under two-and three-part tariffs." *Journal of Marketing Research* 49, no. 6 (2012): 882-899.

Bailey, Murray, ed. *Consumer credit quality: underwriting, scoring, fraud prevention and collections*. White Box Publishing, 2004.

Bellotti, Tony, and Jonathan Crook. "Support vector machines for credit scoring and discovery of significant features." *Expert systems with applications* 36.2 (2009): 3302-3308.

Betancourt, Michael. "A general metric for Riemannian manifold Hamiltonian Monte Carlo." In *Geometric science of information*, pp. 327-334. Springer, Berlin, Heidelberg, 2013.

Betancourt, Michael, Simon Byrne, Sam Livingstone, and Mark Girolami. "The geometric foundations of Hamiltonian Monte Carlo." *Bernoulli* 23, no. 4A (2017): 2257-2298.

Braun, Michael, and Jon McAuliffe. "Variational inference for large-scale models of discrete choice." *Journal of the American Statistical Association* 105.489 (2010): 324-335.

Braun, Michael, and Wendy W. Moe. "Online display advertising: Modeling the effects of multiple creatives and individual impression histories." *Marketing Science* 32, no.5 (2013): 753-767.

Bushway, Shawn, Brian D. Johnson, and Lee Ann Slocum. "Is the magic still there? The use of the Heckman two-step correction for selection bias in criminology." *Journal of Quantitative Criminology* 23, no. 2 (2007): 151-178.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. "Stan: A probabilistic programming language." *Journal of statistical software* 76, no. 1 (2017).

Chen, Mu-Chen, and Shih-Hsien Huang. "Credit scoring and rejected instances reassigning through evolutionary computation techniques." *Expert Systems with Applications* 24.4 (2003): 433-441.

Danaher, Peter J. "Optimal pricing of new subscription services: Analysis of a market experiment." *Marketing Science* 21, no. 2 (2002): 119-138.

Dew, Ryan, and Asim Ansari. "Bayesian Nonparametric Customer Base Analysis with Model-Based Visualizations." *Marketing Science* 37.2 (2018): 216-235.

Dubin, Jeffrey A., and Douglas Rivers. "Selection bias in linear regression, logit and probit models." *Sociological Methods & Research* 18, no. 2-3 (1989): 360-390.

Duchi, John. "Derivations for linear algebra and optimization." *Berkeley, California* 3 (2007).

Dube, Jean-Pierre, Günter J. Hitsch, and Peter E. Rossi. "State dependence and alternative explanations for consumer inertia." *The RAND Journal of Economics* 41.3 (2010): 417-445.

Dzyabura, Daria, and John R. Hauser. "Active machine learning for consideration heuristics." *Marketing Science* 30.5 (2011): 801-819.

Fader, Peter S., Bruce GS Hardie, and Ka Lok Lee. ""Counting your customers" the easy way: An alternative to the Pareto/NBD model." *Marketing science* 24.2 (2005): 275-284.

Fader, Peter S., Bruce GS Hardie, and Jen Shang. "Customer-base analysis in a discrete-time noncontractual setting." *Marketing Science* 29.6 (2010): 1086-1108.

Feinberg, Fred M., Linda Court Salisbury, and Yuanping Ying. "When Random Assignment Is Not Enough: Accounting for Item Selectivity in Experimental Research." *Marketing Science* 35, no. 6 (2016): 976-994.

Feit, Eleanor McDonnell, Mark A. Beltramo, and Fred M. Feinberg. "Reality check: Combining choice experiments with market data to estimate the importance of product attributes." *Management Science* 56.5 (2010): 785-800.

Feit, E.M. and Bradlow, E.T., *Data Fusion*, Handbook of Marketing Research (2016).

Feit, Eleanor McDonnell, et al. "Fusing aggregate and disaggregate data with an application to multiplatform media consumption." *Journal of Marketing Research* 50.3 (2013): 348-364.

Florez-Lopez, R. "Effects of missing data in credit risk scoring. A comparative analysis of methods to achieve robustness in the absence of sufficient data." *Journal of the Operational Research Society* 61.3 (2010): 486-501.

Gal, Y., Van Der Wilk, M., & Rasmussen, C. E. (2014). Distributed variational inference in sparse Gaussian process regression and latent variable models. In *Advances in Neural Information Processing Systems* (pp. 3257-3265).

Gelman, Andrew, and Donald B. Rubin. "Inference from iterative simulation using multiple sequences." *Statistical science* (1992): 457-472.

Gilula, Zvi, Robert E. McCulloch, and Peter E. Rossi. "A direct approach to data fusion." *Journal of Marketing Research* 43.1 (2006): 73-83.

Gilula, Zvi, and Robert McCulloch. "Multi level categorical data fusion using partially fused data." *Quantitative Marketing and Economics* 11.3 (2013): 353-377.

Girolami, Mark, and Ben Calderhead. "Riemann manifold langevin and hamiltonian monte carlo methods." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73, no. 2 (2011): 123-214.

Goettler, Ronald L., and Karen Clay. "Tariff choice with consumer learning and switching costs." *Journal of Marketing research* 48, no. 4 (2011): 633-652.

Golob, Thomas F., and Amelia C. Regan. "Trucking industry adoption of information technology: a multivariate discrete choice model." *Transportation Research Part C: Emerging Technologies* 10, no. 3 (2002): 205-228.

Gu, Zheyin, and Sha Yang. "Quantity-discount-dependent consumer preferences and competitive nonlinear pricing." *Journal of Marketing Research* 47, no. 6 (2010): 1100-1113.

Guadagni, Peter M., and John DC Little. "A logit model of brand choice calibrated on scanner data." *Marketing Science* 2, no. 3 (1983): 203-238.

Grubb, Michael D., and Matthew Osborne. "Cellular service demand: Biased beliefs, learning, and bill shock." *American Economic Review* 105, no. 1 (2015): 234-71.

Hand, David J., and S. D. Jacka. "Consumer credit and statistics." *Statistics in finance* 69 (1998): 81.

Hanssens, Dominique M., Peter SH Leeflang, and Dick R. Wittink. "Market response models and marketing practice." *Applied Stochastic Models in Business and Industry* 21.4-5 (2005): 423-434.

Heckman, James J. "Sample selection bias as a specification error (with an application to the estimation of labor supply functions)." *Econometrica* 47, no. 1 (1979):153–161.

Heckman, James J. "Varieties of selection bias." *The American Economic Review* 80, no. 2 (1990): 313-318.

Hensman, James, Alexander Matthews, and Zoubin Ghahramani. "Scalable variational Gaussian process classification." (2015).

Hilscher, Jens, and Mungo Wilson. "Credit ratings and credit risk: Is one measure enough?." Management science 63.10 (2016): 3414-3437.

Hoffman, Matthew D., et al. "Stochastic variational inference." *The Journal of Machine Learning Research* 14.1 (2013): 1303-1347.

Iyengar, Raghuram, Asim Ansari, and Sunil Gupta. "A model of consumer learning for service quality and usage." *Journal of Marketing Research* 44, no. 4 (2007): 529-544.

Iyengar, Raghuram, and Sunil Gupta. "16 Nonlinear pricing." *Handbook of Pricing Research in Marketing* (2009): 355.

Jordan, Michael I., et al. "An introduction to variational methods for graphical models." *Learning in graphical models*. Springer, Dordrecht, 1998. 105-161.

Kamakura, Wagner A., and Gary J. Russell. "A probabilistic choice model for market segmentation and elasticity structure." *Journal of marketing research* 26.4 (1989): 379-390.

Kamakura, Wagner A., and Michel Wedel. "Statistical data fusion for cross-tabulation." *Journal of Marketing Research*34.4 (1997): 485-498.

Khan, Romana J., and Dipak C. Jain. "An empirical analysis of price discrimination mechanisms and retailer profitability." *Journal of Marketing Research* 42, no. 4 (2005): 516-524.

Kumar, Vineet. "Making 'Freemium' Work: Many Start-ups Fail to Recognize the Challenges of This Popular Business Model." *Harvard Business Review* 92, no. 5 (May 2014): 27–29.

Lambrecht, Anja, Katja Seim, and Catherine Tucker. "Stuck in the adoption funnel: The effect of interruptions in the adoption process on usage." *Marketing Science* 30, no. 2 (2011): 355-367.

Lambrecht, Anja, Katja Seim, Naufel Vilcassim, Amar Cheema, Yuxin Chen, Gregory S. Crawford, Kartik Hosanagar et al. "Price discrimination in service industries." *Marketing Letters* 23, no. 2 (2012): 423-438.

Lambrecht, Anja, and Bernd Skiera. "Paying too much and being happy about it: Existence, causes, and consequences of tariff-choice biases." *Journal of marketing Research* 43, no. 2 (2006): 212-223.

Lee, Timothy H., and Jung Sung-Chang. "Forecasting creditworthiness: Logistic vs. artificial neural net." *The Journal of Business Forecasting* 18.4 (1999): 28.

Levitt, Steven D., John A. List, Susanne Neckermann, and David Nelson. "Quantity discounts on a virtual good: The results of a massive pricing experiment at King Digital Entertainment." *Proceedings of the National Academy of Sciences* 113, no. 27 (2016): 7323-7328.

Levitt, Steven D., and John A. List. "Field experiments in economics: The past, the present, and the future." *European Economic Review* 53, no. 1 (2009): 1-18.

Lim, Michael K., and So Young Sohn. "Cluster-based dynamic scoring model." *Expert Systems with Applications* 32.2 (2007): 427-431.

Little, Roderick JA, and Donald B. Rubin. "Bayes and multiple imputation." *Statistical analysis with missing data* (2002): 200-220.

Kai, Li, and Nagpurnanand R. Prabhala. "Self-selection models in corporate finance." *Handbook of empirical corporate finance*. Elsevier, 2007. 37-86.

Little, Roderick JA, and Donald B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2012, 2nd ed. (John Wiley & Sons, Hoboken, NJ).

Manchanda, Puneet, Grant Packard, and Adithya Pattabhiramaiah. "Social dollars: The economic impact of customer participation in a firm-sponsored online customer community." *Marketing Science* 34, no. 3 (2015): 367-387.

Mason, Alexina, Sylvia Richardson, and Nicky Best. "Two-pronged strategy for using DIC to compare selection models with non-ignorable missing responses." *Bayesian Analysis* 7, no. 1 (2012): 109-146.

McCulloch, Robert, and Peter E. Rossi. "An exact likelihood analysis of the multinomial probit model." *Journal of Econometrics* 64, no. 1 (1994): 207-240.

McFadden, Daniel. "Conditional logit analysis of qualitative choice behavior." (1973).

McManus, Brian. "Nonlinear pricing in an oligopoly market: The case of specialty coffee." *The RAND Journal of Economics* 38, no. 2 (2007): 512-532.

Mussa, Michael, and Sherwin Rosen. "Monopoly and product quality." *Journal of Economic theory* 18, no. 2 (1978): 301-317.

Neal, Radford M. "MCMC using Hamiltonian dynamics." *Handbook of Markov Chain Monte Carlo* 2, no. 11 (2011).

Niculescu, Marius F., and Dong Jun Wu. "Economics of free under perpetual licensing: Implications for the software industry." *Information Systems Research* 25.1 (2014): 173-199.

Ong, Chorng-Shyong, Jih-Jeng Huang, and Gwo-Hshiung Tzeng. "Building credit scoring models using genetic programming." *Expert Systems with Applications* 29.1 (2005): 41-47.

Orgler, Yair E. *Evaluation of bank consumer loans with credit scoring models.* Tel-Aviv University, Department of Envirnonmental Sciences, 1971.

Pakman, Ari, and Liam Paninski. "Exact hamiltonian monte carlo for truncated multivariate gaussians." *Journal of Computational and Graphical Statistics* 23, no. 2 (2014): 518-542.

Petrin, Amil, and Kenneth Train. "A control function approach to endogeneity in consumer choice models." *Journal of marketing research* 47, no. 1 (2010): 3-13.

Polson, Nicholas G., James G. Scott, and Jesse Windle. "Bayesian inference for logistic models using Pólya–Gamma latent variables." *Journal of the American statistical Association* 108.504 (2013): 1339-1349.

Puranam, Dinesh, Vishal Narayan, and Vrinda Kadiyali. "The effect of calorie posting regulation on consumer opinion: a flexible Latent Dirichlet Allocation model with informative priors." *Marketing Science* 36.5 (2017): 726-746.

Qian, Yi, and Hui Xie. "No customer left behind: A distribution-free bayesian approach to accounting for missing xs in marketing models." *Marketing Science* 30.4 (2011): 717-736.

Qian, Yi, and Hui Xie. "Which brand purchasers are lost to counterfeiters? An application of new data fusion approaches." *Marketing Science* 33.3 (2013): 437-448.

Raghunathan, Trivellore E., and James E. Grizzle. "A split questionnaire survey design." *Journal of the American Statistical Association* 90.429 (1995): 54-63.

Rao, Vithala R., ed. *Handbook of pricing research in marketing.* Edward Elgar Publishing, 2009.

Ratchford, Brian T. "Online pricing: review and directions for research." *Journal of Interactive Marketing* 23.1 (2009): 82-90.

Rässler, Susanne, Florian Koller, and Christine Mäenpää. *A split questionnaire survey design applied to german media and consumer surveys.* No. 42b/2002. Diskussionspapiere//Friedrich-Alexander-Universität Erlangen-Nürnberg, Lehrstuhl für Statistik und Ökonometrie, 2002.

Reiley, David H. "Field experiments on the effects of reserve prices in auctions: More magic on the internet." *The RAND Journal of Economics* 37, no. 1 (2006): 195-211.

Rubin, Donald B. "The use of matched sampling and regression adjustment to remove bias in observational studies." *Biometrics* (1973): 185-203.

Ruiz, Francisco JR, et al. "Augment and reduce: Stochastic inference for large categorical distributions." *arXiv preprint arXiv:1802.04220* (2018).

Sarlija, Natasa, Mirta Bensic, and Zoran Bohacek. "Multinomial model in consumer credit scoring." 10th International Conference on Operational Research (10; 2004). 2004.

Schmittlein, David C., Donald G. Morrison, and Richard Colombo. "Counting your customers: Who-are they and what will they do next?." *Management science* 33.1 (1987): 1-24.

Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64, no. 4 (2002): 583-639.

Snelson, Edward, and Zoubin Ghahramani. "Sparse Gaussian processes using pseudo-inputs." *Advances in neural information processing systems.* 2006.

Sullivan, A. C. "Consumer finance." EI Altman, *Financial Handbook* (9.3-9.27), New York: John Wiley & Sons (1981).

Sustersic, Maja, Dusan Mramor, and Jure Zupan. "Consumer credit scoring models with limited data." *Expert Systems with Applications* 36.3 (2009): 4736-4744.

Swait, Joffre, and Jordan Louviere. "The role of the scale parameter in the estimation and comparison of multinomial logit models." *Journal of marketing research* 30.3 (1993): 305-314.

Tank, Alex, Nicholas Foti, and Emily Fox. "Streaming variational inference for Bayesian nonparametric mixture models." *Artificial Intelligence and Statistics*. 2015.

Tanner, Martin A., and Wing Hung Wong. "The calculation of posterior distributions by data augmentation." *Journal of the American Statistical Association* 82, no. 398 (1987): 528-540.

Thomas, Lyn C., David B. Edelman, and Jonathan N. Crook. *Credit scoring and its applications*. Society for industrial and Applied Mathematics, 2002.

Thurstone, Louis L. "A law of comparative judgment." Psychological review 34.4 (1927): 273.

Titsias, Michalis. "One-vs-each approximation to softmax for scalable estimation of probabilities." *Advances in Neural Information Processing Systems*. 2016.

Train, Kenneth E., Moshe Ben-Akiva, and Terry Atherton. "Consumption patterns and self-selecting tariffs." *The Review of Economics and Statistics* (1989): 62-73.

Train, Kenneth E., Daniel L. McFadden, and Moshe Ben-Akiva. "The demand for local telephone service: A fully discrete model of residential calling patterns and service choices." *The RAND Journal of Economics* (1987): 109-123.

Tversky, Amos. "Elimination by aspects: A theory of choice." *Psychological Review* 79, no. 4 (1972): 281.

Wachtel, Stephan, and Thomas Otter. "Successive sample selection and its relevance for management decisions." *Marketing Science* 32.1 (2013): 170-185.

Wei, Yanhao, et al. "Credit scoring with social network data." Marketing Science 35.2 (2015): 234-258.

Wilson, Robert B. *Nonlinear pricing*. Oxford University Press on Demand, 1993.

Winship, Christopher, and Robert D. Mare. "Models for sample selection bias." *Annual review of sociology* 18, no. 1 (1992): 327-350.

Winer, Russell S., and Scott A. Neslin, eds. *The history of marketing science.* New York, NY: World Scientific, 2014.

Wenzel, Florian, et al. "Efficient Gaussian process classification using Pòlya-Gamma data augmentation." *arXiv preprint arXiv:1802.06383* (2018).

Wolk, Agnieszka, and Bernd Skiera. "Tariff-specific preferences and their influence on price sensitivity." *Business Research* 3, no. 1 (2010): 70-80.

Wu, Yue, Kaifu Zhang, and V. Padmanabhan. "Matchmaker competition and technology provision." *Journal of Marketing Research* 55, no. 3 (2018): 396-413.

Zanutto, Elaine L., and Eric T. Bradlow. "Data pruning in consumer choice models." *Quantitative Marketing and Economics* 4, no. 3 (2006): 267-287.

Zhang, Xiao, W. John Boscardin, and Thomas R. Belin. "Bayesian analysis of multivariate nominal measures using multivariate multinomial probit models." *Computational statistics & data analysis* 52, no. 7 (2008): 3697-3708.