CARNEGIE MELLON UNIVERSITY

DOCTORAL DISSERTATION

Essays on Bayesian Machine Learning in Marketing

Samuel LEVY

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Marketing

Tepper School of Business
Carnegie Mellon University

April, 2024

Dissertation Committee

Alan Montgomery (Chair)
Carnegie Mellon University
Tepper School of Business

Asim Ansari
Columbia University
Graduate School of Business

Tim Derdenger
Carnegie Mellon University
Tepper School of Business

Fred Feinberg
University of Michigan
Ross School of Business

Joy Lu
Carnegie Mellon University
Tepper School of Business

Kannan Srinivasan
Carnegie Mellon University
Tepper School of Business

Abstract

Chapter 1, titled *"Privacy Preserving Data Fusion"*, and joint work with Longxiu Tian and Dana Turjeman, tackles the complex problem of merging multiple datasets while ensuring user privacy. This paper introduces a privacy-preserving data fusion methodology that adheres to the principles of differential privacy, leveraging variational autoencoders and normalizing flows to create a robust, nonparametric, Bayesian generative modeling framework. This methodology notably accounts for missingness in each dataset, correcting for sample selection and negating the requirement for identical users across datasets when learning the joint data generating process. Through a series of simulations and an applied case involving a novel large-scale customer satisfaction survey and CRM database from a leading U.S. telecom carrier, we demonstrate the potential of this privacy-preserving methodology for robust data fusion, providing insights into customer satisfaction and churn propensity without compromising privacy.

Chapter 2, titled *"Understanding Consumer Expenditure Through Gaussian Process Choice Models"*, is joint work with Alan Montgomery. This chapter challenges the rigid structural assumptions of traditional choice models that define expenditure elasticity and the restrictive utility functional forms these models often impose. By introducing Gaussian process priors on utility functions, we provide a flexible, utility - based model for understanding expenditure - driven changes in consumer choices. We demonstrate that relaxing the functional form on the outside good within the framework of constrained utility maximization leads to more flexible substitution patterns. This has implications for understanding preference for variety and quality. This methodological advance enables the model to capture non-linear rates of satiation and precise baseline preferences—details that traditional non - homothetic (i.e., expenditure-variant preferences) parametric models often overlook due to their assumptions of a given utility functional form. Through its automatic detection of non-linear consumption patterns from the data, the model provides more flexible statistical inference, offering valuable theoretical and practical insights for improved pricing decisions.

Chapter 3, titled *"Digital Twins: A Generative Approach for Counterfactual Customer Analytics"*, proposes an innovative methodology to optimize customer surveys in a competitive landscape. Leveraging a unique dataset of quarterly cross-sectional survey responses from major U.S. telecommunications providers from 2020 to 2022, this paper introduces the concept of 'Digital Marketing Twins.' These are generative models of customer preferences that provide counterfactual responses under different scenarios. Here, the concept of “generative model” means that I explicitly give the sequence of steps describing how the data were created, i.e., the data generating process, including unknown model parameters. The methodology uses a novel deep generative and probabilistic latent factor model, which captures individual - level brand affinity for each brand and time period, accounting for observed heterogeneity and firm-side factors. Utilizing Bayesian optimization, the model offers individual-level marketing action recommendations. It shows promising results in identifying marketing actions most likely to increase customer satisfaction, offering a “path of least resistance” at the individual level.

Acknowledgements

I am profoundly grateful to my dissertation chair Alan Montgomery for his unparalleled guidance and mentoring throughout my PhD journey. Alan has been instrumental in shaping my research topics, sparking my interests in methodological work within the field of marketing, and teaching me the virtue of patience. Alan has also taught me Bayesian statistics for marketing (my favorite class!), tremendously helped with data acquisition, and generously funded my research. But most importantly, he helped me become an independent researcher. His wisdom and support have been a cornerstone of my academic development. Alan, thank you so much for being such a great advisor.

I extend my heartfelt thanks to my committee members: Fred Feinberg, Asim Ansari, Tim Derdenger, Joy Lu, and Kannan Srinivasan. Joy's door was always open to chat about research and life, and I am really grateful to have her on my committee. Fred has been incredibly helpful before and during my job market, and provided great mentorship. *Mercy buckets*, Fred!

Special acknowledgment goes to my co-author and friend Longxiu Tian for his mentoring, essential support in data acquisition and preparation for the job market. Longxiu acted like an academic big brother and I am so grateful for that. I also thank my co-author and friend Dana Turjeman for her continuous guidance. It's been great working with you. Dokyun Lee deserves a special mention for igniting my interest in Bayesian machine learning in marketing. I thank DK for his encouragement in the early stages of my PhD and his ability to fill me with ideas and optimism. DK also introduced me to the Bayesian computation reading group where I met Longxiu and many others. I would like to also acknowledge and thank Nish and Emmanuel for being such wonderful industry partners.

My sincere appreciation goes to Olivier Rubel for over a decade of continuous mentoring and friendship. Since the days of my *agrégation* in France, Olivier has been a constant source of guidance and support – when I had something to celebrate, or when things were not going well, Olivier was always here through text, phone calls, and in-person meetings in France and in California. His mentorship has profoundly shaped my academic path. Olivier – when do we start writing this paper together?

I have been blessed with the friendship and camaraderie of many wonderful people during my PhD: Martin Michelini, Zahra Ebrahimi, Jaepil Lee and Joohyun

Lee, Behnam Mohammadi and Elisabeth Estes, Thomas Devine, Kevin Mott and Sae-Seul Park, Lavender Yang and Qiaochu Wang, Serim Hwang, Julie Wang, Flora Feng, Yuan Yuan, Angela Xiao, Jingwei Dai, Savannah Tang, Jason Gates, Zhaoqi Cheng, David Li, Jaymo Kim, Liying Qiu, Anthony Karahalios, Daniel de Roux, Genaro Basulto, Nicholas Hoffman, Kerry Zhang, Minhyuk Nam, Arnav Sood, Orhun Gun, Neha Neha, Sebastian Vasquez, Macarena Navarro, Allen Brown, Nate Fulham, Jenny Oh, Frederic and Sumeyye Gmeiner, my job market buddy Jimin Nam, my buddy Clément Boullé and his wife Parama Gosh, my former roommate Ryan Kavanagh, Léa Sonza, and my oldest friend to date, Arnaud Buzzi. Your friendship, support, and shared moments have made this journey not only bearable but enjoyable.

I am grateful to Laila Lee and Lawrence Rapp for their administrative support. Your assistance has been invaluable in navigating the complexities of the PhD program.

I owe an immense debt of gratitude to my family: my mother Isabelle, my father Michel, my sisters Lorenza and Nathalie, and my brother Jonathan. Your unwavering moral support and belief in my abilities have been my strength. Thank you for your love and encouragement. *Je vous embrasse très fort.*

Finally, a special thank you to my fiancée, Nur Yıldırım. Meeting you on the first day of orientation in the PhD program and growing together both personally and professionally has been one of the greatest joys of my life. Becoming faculty together marks the beginning of our exciting new chapter. This dissertation, and all my achievements, would not have been possible without your love, support, and partnership. I am incredibly lucky to have you by my side.

Contents

List of Figures

List of Tables

Chapter 1

Privacy Preserving Data Fusion

joint with Longxiu Tian and Dana Turjeman

Data fusion combines multiple datasets to make inferences that are more accurate, generalizable, and useful than those made with any single dataset alone. However, data fusion poses a privacy hazard due to the risk of revealing user identities. We propose a privacy preserving data fusion (PPDF) methodology intended to preserve user-level anonymity while allowing for a robust and expressive data fusion process. PPDF is based on variational autoencoders and normalizing flows, together enabling a highly expressive, nonparametric, Bayesian, generative modeling framework, estimated in adherence to differential privacy – the state-of-the-art theory for privacy preservation. PPDF does not require the same users to appear across datasets when learning the joint data generating process and explicitly accounts for missingness in each dataset to correct for sample selection. Moreover, PPDF is model-agnostic: it allows for downstream inferences to be made on the fused data without the analyst needing to specify a discriminative model or likelihood *a priori*. We undertake a series of simulations to showcase the quality of our proposed methodology. Then, we fuse a large-scale customer satisfaction survey to the customer relationship management (CRM) database from a leading U.S. telecom carrier. The resulting fusion yields the joint distribution between survey satisfaction outcomes and CRM engagement metrics at the customer level, including the likelihood of leaving the company's services. Highlighting the importance of correcting selection bias, we illustrate the divergence between the observed survey responses vs. the imputed distribution on the customer base. Managerially, we find a negative, nonlinear relationship between satisfaction and future account termination across the telecom carrier's customers, which can aid in segmentation, targeting, and proactive churn management. Overall, PPDF will substantially reduce the risk of compromising privacy and anonymity when fusing different datasets.

1.1 Introduction

Data fusion consists of the combination or linkage of multiple data sources to make inferences that are more accurate, generalizable, and useful than those made with any single dataset alone. Data fusion has been applied at leading technology firms, including Facebook (Theo Ryffel et al., 2018), Microsoft (Zheng, 2015), and Google (Papernot, 2019). Managers use data fusion to learn about common buying behaviors, customer preferences, and prospective needs, from physically and conceptually distinct datasets. For example, data fusion assists in learning about customer needs through the fusion of choice surveys and eventual purchase data (Eleanor McDonnell Feit, Beltramo, and Feinberg, 2010), or in making more accurate predictions of potential market share, through the fusion of data on both customers and the general population (McCarthy and Oblander, 2021). Despite the prevalence and advantages of data fusion, however, whenever data fusion involves any form of customer-level data[1], the technique poses a privacy hazard of identifying individuals. For example, Sweeney, 1997, Narayanan and Shmatikov, 2008, and S. Li et al., 2022 show that a combination of datasets might reveal individuals' sensitive and identifiable information, in a process referred to as "linkage attacks." The data to be fused, even if anonymous or de-identified, might be re-identified when fused, therefore risking the privacy of individuals in either dataset.

To reduce the risks of user identification, while allowing for the advantages of data fusion, we develop a Privacy Preserving Data Fusion (PPDF) methodology, based on differential privacy (DP) (Dwork, Kenthapadi, et al., 2006), variational autoencoders (VAE) (Kingma and Welling, 2013), and normalizing flows (NF) (D. Rezende and Mohamed, 2015). It fuses two or more datasets nonparametrically and generatively, and implements differential privacy, a state-of-the-art framework and methodology to assure privacy preservation. PPDF will allow organizations that handle individual - level data to substantially reduce, or even eliminate, the risk of compromising privacy and anonymity.

Beyond being a critical pillar of customer privacy, assurance of anonymity has been shown to increase survey response rate and honesty (Bradburn et al., 1979). Firms often hold additional data sources such as customer relationship management (CRM) or behavioral data. The data from different sources can be fused with survey data, in order to gather insights on the customer base.

Hence, developing privacy preserving methodologies is becoming a prominent need. With privacy in mind, the overarching goal is to find the balance between privacy and the great advantages provided by data and data - driven decision making. Customers are increasingly aware and demanding of privacy (Lin, 2022). Consequently, marketing efforts - previously focused on products' abilities - are now shifting to protecting customers' privacy as part of a broader push towards customer - centricity in marketing.

We exemplify the use of PPDF and illustrate its advantages by conducting analyses on several different domains and datasets:

[1] The nouns user, customer, individual, person, and consumer will be used interchangeably to describe people whose information, some of which may be private and/or identifiable, is held by companies.

1. We use the well-known MNIST image dataset of hand-written digits to visually intuit the methodology and the advantages of each building-block, including: data fusion accuracy, scalability, and the trade-off between privacy and accuracy. Details are provided in Section 1.1.

2. A simulation using a survey of 5.5K respondents that is split and then re-fused, to show both the trade-off between privacy and accuracy, as above, and also the sensitivity to other tuning parameters of PPDF. More on this in Section 1.4.2.

3. Our main illustrative application will be in fusing CRM data from a leading U.S. telecommunications company (hereafter “telecom carrier" or “firm") to a large-scale customer satisfaction survey of its customer base conducted by an external surveying company. Crucially, to ascertain the accuracy of PPDF, the telecom carrier further conducted an identical internal survey in which we know the user-id in both the survey and the CRM data[2]. This exercise allows us to again show the trade-off between accuracy and privacy and to also illustrate the managerial value of PPDF. Details of these analyses are provided in Section 1.5.

In the application of PPDF to the telecom carrier's CRM and survey, the fusion allows us to assess, among other things, the relationship between self-reported customer satisfaction metrics from the survey and the likelihood to churn from the company's services across the full customer base, yet without revealing customers' identities as being among the survey's respondents. Of note, in anonymous survey settings, firms are limited to analyzing solely the sample distribution of the responses, which can suffer from selection bias. Of particular interest to the firm's managers, our data fusion reveals that the relationship between predicted Likelihood to Recommend (LTR) and churn from the full customer base has a reverse hockey stick-shaped relationship. Whereas one would expect a straightforward negative correlation between LTR and churn, instead, we see that those with the highest predicted LTR (10/10) are more likely to churn than those with a predicted LTR of 6/10. Moreover, this exercise represents the first-ever individual-level ‘scoring' of the firm's existing customers with their imputed LTR value. More generally, we quantify the distributional divergence between survey respondents and the full customer base. This highlights the well-known phenomenon of selection bias in survey outcomes and how the proposed data fusion assists in correcting such selectivity to achieve more accurate managerial insights and decision-making.

Our work makes several contributions: First, we propose a new, nonparametric and scalable data fusion methodology to generate customer insights from disparate

[2] All CRM data and surveys were cleaned of personally identifiable information, including but not limited to names, phone numbers, street addresses, and emails. Moreover, account numbers were re-enumerated using a random index (hereafter “user ID"), prior to data-sharing, as stipulated in the Non-Disclosure Agreement (NDA) signed by the telecom carrier and the authors.

data sources. We demonstrate the quality of the data fusion in multiple data types and contexts, in terms of both accuracy and posterior predictive ability. Second, our data fusion methodology has a built-in privacy tuning mechanism that can be tuned by the data holders. We quantitatively show the trade-off between privacy and accuracy. Third, we present an application where we successfully fuse anonymous survey data and CRM data under privacy constraints. This application allows us not only to exemplify the method, but also to predict if a customer is at risk of leaving the company. This adds another building block in the research on the relationship between customer satisfaction and engagement, but is only the tip of the iceberg, and solely an example, of explorations that will become feasible with the proposed PPDF methodology.

To the best of our knowledge, PPDF is the first methodology that enables researchers and managers who handle sensitive data to securely fuse datasets and gather inferences that cannot be made with each dataset alone. Our data fusion methodology comes with known and principled privacy guarantees (Dwork, Kenthapadi, et al., 2006; Abadi et al., 2016). Data holders can use the methodology in collaboration with other entities, without revealing data and without risking the privacy of individuals present in them. Our methodology does not require the same users to appear in both datasets and explicitly accounts for missingness (e.g., sample selection bias). Moreover, PPDF is the first fully generative data fusion technique; that is, it is model-agnostic as the analyst need not specify the discriminative model/analysis prior to fusing data.

As a high level description of our methodology, PPDF consists of three core components in its specification and estimation: differential privacy (DP), variational autoencoders (VAE), and normalizing flows (NF). Differential privacy (Dwork, McSherry, et al., 2006) introduces the concept of a “privacy budget” where researchers can set an adequate level of privacy at the price of lower accuracy. The main assumption is that privacy of users is preserved when one cannot identify any particular user as being included in the dataset. In probabilistic terms, there is a bounded probability for any individual to be revealed. DP is a state-of-the-art technique among privacy-preserving methodologies. More specifically, we implement a differentially private, doubly-stochastic variational inference algorithm, in its (ε, δ) form (Jälkö, Dikmen, and Honkela, 2017), which is derived from differentially private version of stochastic gradient descent (Abadi et al., 2016). This implementation is realized in the estimation (i.e., training) of PPDF’s variational autoencoders. VAEs are an attractive approach to data fusion because they form a highly scalable and probabilistic generative model designed to learn a concise data generating process in the form of a joint distribution across features. We extend VAEs to also learn across customer-level datasets, with the challenge that they do not have deterministic linkages between observations (i.e., due to anonymity in the survey). As a result, PPDF fuses two or more datasets governed by differential privacy and allows analysts to substantially reduce, or even eliminate, the risk of compromising customers' privacy and anonymity. One caveat of VAEs is that they tend to be limited in their ability to capture data generating processes. We overcome this challenge by allowing the generative distribution to be nonparametrically learned from the data itself, via a normalizing flow architecture. Normalizing flows (D. Rezende and Mohamed, 2015; Papamakarios et al., 2021) refer to a sequence (flow) of non-linear, bijective (volume-preserving, invertible) transformations on probability density functions. A key computational advantage in the context of large-scale data fusions, such as those in our application, is that these transformations are composed from simple distributions (e.g., the standard Normal distribution) and through the diffusion of flows beget a far richer and more accurate representation of the underlying data generating process while retaining the inherent scalability of VAEs.

The rest of the chapter is organized as follows: In Section 1.2, we review the literature on data fusion and privacy preserving methodologies. Our methodology is detailed in Section 1.3. Specifically, we detail the variational autoencoders in Subsection 1.3.1, specify the privacy enhancement in Subsection 1.3.3, and discuss the types of missing data PPDF can handle in Subsection 1.3.4. In Section 1.4, we show our proposed methodology's abilities using several simulations. In Section 1.5, we describe the data used and the results generated herein. Finally, we conclude with a brief summary and a discussion on further applications and future directions in Section 1.6.

1.2 Literature Review

Work in the domain of data fusion and record linkage can be traced back to Dunn, 1946, where multiple population datasets were combined. Contemporary record linkage and more complex forms of data fusion have been used in multiple fields, yielding results in economics (Berry, Levinsohn, and Pakes, 2004), geography (Liu et al., 2020; Dias et al., 2019), and health (Dautov, Distefano, and Buyya, 2019), among others. In the marketing domain, they have been used to handle missing data in surveys (Bradlow and Zaslavsky, 1999), enrich parameter estimates and preference predictability (Swait and Andrews, 2003), estimate product purchasing and media-watching (Gilula, McCulloch, and Rossi, 2006), combine choice experiments with CRM data (Eleanor McDonnell Feit, Beltramo, and Feinberg, 2010), detect heavy and light users in multiple media platforms (Eleanor McDonnell Feit, P. Wang, et al., 2013), predict users' choices based on contextual data from their phones (Unger et al., 2018), and predict market share (McCarthy and Oblander, 2021), among other use cases.

The proposed methodology, similarly to other data fusion methodologies in marketing, is intended to enhance user- and customer-level data by fusing them with other datasets. However, much prior work in this area has focused on fusing detailed individual (disaggregate) data with data that are aggregated across customers (e.g., Eleanor McDonnell Feit, P. Wang, et al., 2013; McCarthy and Oblander, 2021). For such aggregate-disaggregate uses, privacy is less of a concern because linkage attacks are unlikely to occur. This is mostly because data that are aggregated across customers cannot usually shed light on identities of the people who are in the disaggregate (individual-level) data. PPDF, on the other hand, can fuse data from different sources while protecting individuals' privacy, whether the data are aggregated or not.

PPDF methodology does not require that the same customers appear in both datasets to make inferences on the joint data. Fusion occurs based on the joint distribution of the shared and unique variables, and therefore, under standard assumptions of missingness in the data (selection bias being a specific example of missingness in data), to be further described in Section 1.3.4, it recovers one dataset's missingness from additional variation made available in the other dataset. In contrast to prior approaches, an explicit selectivity correction need not be specified in the model (notably, there is no underlying model specification in our applications). Instead, PPDF recovers the missingness in a nonparametric manner, inspired by advances in Bayesian canonical correlation analysis (Klami, Virtanen, and Kaski, 2013; Chandar et al., 2016), and treats each dataset as if it were a random sample from a multivariate random distribution we wish to encode and fuse. As a generative model, PPDF's imputation of missing values takes into account the joint distribution across variables and datasets. For example, if survey respondents' characteristics are different from those of the wider customer base, then sampling from the generative model for the wider customer base should lead to a different LTR distribution. This mechanism automatically eliminates any lack of representativeness arising from self-selection of survey respondents. Therefore, managers who wish to learn from a survey jointly with CRM data, or from other datasets that inherently entail sample selection or other missingness, have more opportunities to do so. More importantly, distinguished from other data fusion methods in marketing, to the best of our knowledge, PPDF is the first fully generative data fusion technique; that is, it is model-agnostic as the analyst need not specify the discriminative model/analysis prior to fusing data. Simultaneously, the generative distribution that is learned on the fused data allows for uncertainty propagation to any downstream inference if the analyst so chooses.

Until now, we have discussed advances in data fusion. With the great advances of data fusion comes great progress, but on the other hand comes the risk of identifying individuals. When both datasets jointly include identifiers, data fusion might compromise customers' privacy (within either dataset) and reveal one's preferences or values along with their identifiable information. This has been illustrated by the seminal work of Sweeney, 1997, who relied on demographic data to reveal sensitive health information of public officials in the State of Massachusetts, and by Narayanan and Shmatikov, 2008, who relied on inferred preferences when matching de-identified data from Netflix with publicly available data from IMDb. Lin and Misra, 2022 discuss how firms such as Google, Apple, and Facebook do not allow identity matching across platforms and across devices, to prevent user identification, and show how such protection may lead to an identity fragmentation bias when an external actor aims to measure customer behaviors. The proposed PPDF methodology allows us to learn the joint distribution of both datasets, based on their latent constructs, without compromising anonymity of any user, and potentially alleviates some of the concerns proposed by Lin and Misra, 2022. PPDF complements recent work by Anand and C. Lee, 2023 who develop a deep learning method for data sharing. As opposed to Anand and C. Lee, 2023, PPDF will enable data sharing with the privacy guarantees of differential privacy and will also develop a full generative model to allow for robust data fusion of the datasets.

Therefore, beyond extending the stream of work on data fusion, PPDF methodology also extends the growing stream of privacy preserving methodologies, such as privacy preserving data publication, training, inference, and synthesis (e.g., Fung et al., 2010; Shokri and Shmatikov, 2015; Ping, Stoyanovich, and Howe, 2017; Takagi et al., 2020; Evans et al., 2019; Kaissis et al., 2021; Anand and C. Lee, 2023. See K. Kim and Tanuwidjaja, 2021 for a recent review). Similar to many of the methods mentioned above, our method builds on differential privacy (Dwork, McSherry, et al., 2006), the gold-standard method for privacy preservation, further explained in Subsection 1.3.3. DP relies on mathematical guarantees and allows for a pre-specified "privacy budget" that can be tuned to the desired risk assessment and tolerated accuracy loss.

Other privacy preserving methodologies, such as K-anonymity (Sweeney, 2002; S. Li et al., 2022) (obscuring the data such that every person cannot be distinguished from other $K-1$ people in the dataset) and $\mathscr{L}$-diversity (assuring that each variable has at least $\mathscr{L}$ well-represented values, Machanavajjhala et al., 2007) have been proposed to enable data publication and data synthesis. While such methods may be relevant for datasets with a relatively small number of attributes, they fail to scale to large datasets and might still suffer from various privacy attacks that could reveal identities of the people represented in the data(N. Li, T. Li, and Venkatasubramanian, 2007; Domingo-Ferrer and Torra, 2008). Nevertheless, they have been found suitable for multiple uses, most notably password checkup tools such as "Have I been Pwned" and Google's security checkup (L. Li et al., 2019). Recently, S. Li et al., 2022 extend K-anonymity for longitudinal panel data to also protect against linkage attacks. We rely on differential privacy due to its ability to handle richer datasets and due to the mathematical guarantees and clear tuning parameters it enables. In our simulation exercise, we will demonstrate the privacy preservation vs. accuracy trade-off visually, using the MNIST dataset of hand-written images, as well as using data from another survey, unrelated to the main application, which we split and then fuse back with varying tuning parameters and varying privacy guarantees. Finally, we explore the trade-off between accuracy and privacy in the main application on the telecom carrier data, using an internal calibration survey in which we have full information about users' identities.

In our main application, using the telecom carrier data, we explore how "Likelihood to Recommend" may assist in predicting customers' churn. The general ability to predict churn, and the relationship between LTR and churn, has been a source of debate, both in industry and in academic settings (e.g., Lemmens and Croux, 2006; De Haan, Verhoef, and Wiesel, 2015; Neslin et al., 2006; Ascarza et al., 2018; Lemmens and Gupta, 2020). With PPDF, a company will be better able to predict churn and potentially offer tools to improve the antecedents and outcomes of low customer satisfaction. Of note, this is just a specific illustration of a managerial use of PPDF methodology. The generative framework underlying PPDF ensures that future end-users may craft specific analyses based on their datasets and context, which can extend beyond surveys and CRM data.

1.3 PPDF Methodology

In this section, we first give an overview of the model before decomposing each component and explaining its role in the proposed methodology: variational autoencoders and normalizing flows; learning mechanism of the data fusion; privacy preservation measures and controls.

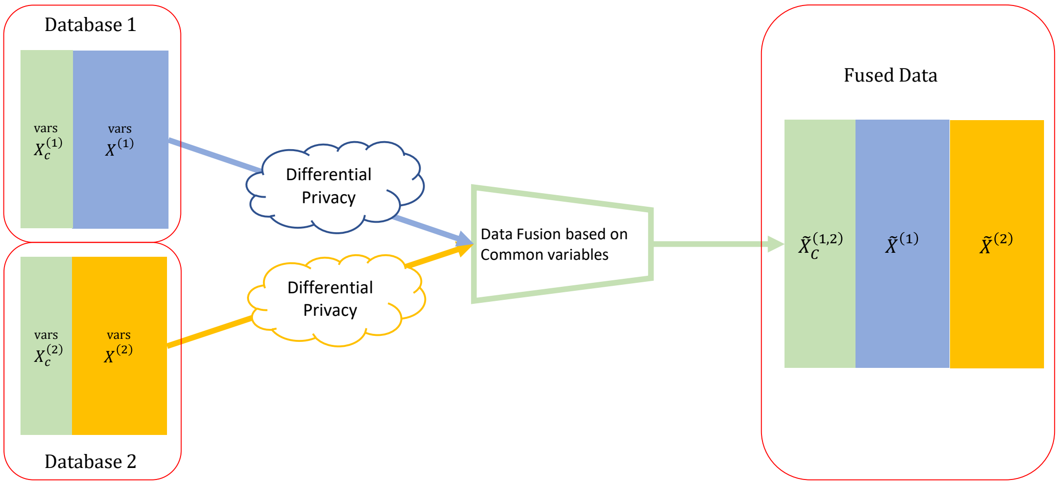


FIGURE 1.1: High level illustration of PPDF of two datasets, each with some common variables $X^{(c)}$. Dataset 1 has variables $X^{(1)}$ and Dataset 2 has variables $X^{(2)}$. The datasets go through differential privacy and are then fused into inferred variables $\tilde{X}^{(1)}$, $\tilde{X}^{(2)}$, and $\tilde{X}_C^{(1,2)}$ based on the population of the respective dataset.

Figure 1.1 provides an overview of PPDF and illustrates the architecture on two datasets[3]: Dataset 1, which is comprised of set of variables $X^{(1)}$ (in the telecom carrier example, such CRM data will include, for example, number of lines for the account, customer tenure, contract status, billing information) and common (shared) variables $X_c^{(1)}$ (e.g., engagement metrics such as purchase of a new phone or connected device, reward redemption behavior, recent visits to a retail location) and Dataset 2, which also includes the same common variables $X_c^{(2)}$, but has unique variables $X^{(2)}$ (e.g., various types of user satisfaction measures, such as LTR, overall satisfaction, and more).

Importantly, while $X_c^{(1)}$ and $X_c^{(2)}$ are common variables in that they have similar structure, they might not be of the same users and do not have to be of the same size. The two instances of these shared variables might not even be drawn from the same distribution. For example, customers that are more extreme in their attitudes towards the brand may be more likely to self-select into responding to a survey, and this attitudinal difference then translates into a different distribution in terms of common variables. Such missingness that is due to selection bias can be overcome with our method as long as there is sufficient ability (i.e., enough information) to recover the joint distribution of those who chose not to respond. We detail our ability to overcome selection bias in Section 1.3.4.

Our goal is to infer the joint distribution of the fused data while reducing privacy risks associated with such linkage of datasets. In our focal context of the telecom carrier's wireless customers, we fuse the two sources of data without revealing which users responded to the external survey. Specifically, we find how the attributes from the CRM database ($X^{(1)}$) *covary* with the response outcomes from the anonymous customer survey ($X^{(2)}$) and explicitly obviate any one-to-one *'matches'* between the two datasets.

In the following subsections, we will explain the building blocks of PPDF – starting from discussing a single dataset's encoder and decoder implemented with a variational autoencoder (VAE), improving it through normalizing flow, making it differentially private. Then, shifting back to discussing two datasets, we will explain how we use the VAE's encoder and decoders as a shared learning mechanism to fuse the two datasets.

1.3.1 Variational Autoencoders (VAEs)

Variational autoencoders have been used widely to capture the generative process of images and other data types. In this subsection, we will describe the variational autoencoders included in PPDF. Figure 1.2 illustrates the basic setup for a VAE that learns a joint representation for a single dataset with two sets of variables, $X_c^{(1)}$ and

[3] In what follows, and for ease of notation, we assume two datasets are to be fused, though this can be generalized.

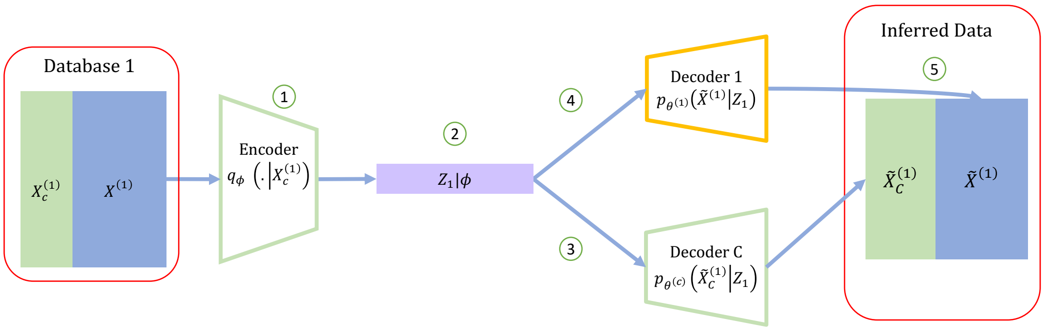

Database 1
$X_c^{(1)}$
$X^{(1)}$
1
Encoder
$q_\phi\left(.|X_c^{(1)}\right)$
2
$Z_1|\phi$
4
3
Decoder 1
$p_{\theta^{(1)}}(\tilde{X}^{(1)}|Z_1)$
Decoder C
$p_{\theta^{(c)}}\left(\tilde{X}_c^{(1)}|Z_1\right)$
Inferred Data
5
$\tilde{X}_C^{(1)}$
$\tilde{X}^{(1)}$

FIGURE 1.2: Illustration of a variational autoencoder (VAE) of a single dataset (without loss of generality, Dataset 1). Two parts of the dataset, variables $X_c^{(1)}$ and $X^{(1)}$, are encoded to the latent variable Z_1 through the function $q(.)$, parameterized by the variational vector of parameters ϕ through a neural network (1). Note that the variational family is a function of the input and the shared variational parameters (i.e., it is amortized). The latent vector Z_1 (2) is then decoded *via* two decoders: Decoder 1 and Decoder C (for “common"), parameterized by θ through a neural network (3,4), to reconstruct the original data (5). The VAE is a stochastic computational graph that simultaneously optimizes the variational parameters ϕ and the model parameters θ.

$X^{(1)}$. The VAE learns the generative model of the dataset, and has two types of components:

- An *encoder* (also known as an inference or recognition model) uses the two variable sets as inputs $X^{(1)}$ and $X_c^{(1)}$ (where the c stands for common variables) and estimates a set of latent representations $q_\phi(.|X_c^{(1)})$ with inference parameters ϕ that capture the data generating process into latent representation Z_1.

- Two *decoders* (also known as an amortized inference or generative model) take Z_1 as input and estimate two conditionally independent models $p_\theta^{(1)}(\tilde{X}^{(1)}|Z_1)$ and $p_\theta^{(c)}(\tilde{X}^{(c)}|Z_1)$ used to reconstruct the original data with set of parameters $\theta \equiv (\theta^{(1)}, \theta^{(c)})$.

The difference between the original data $X \equiv (X^{(1)}, X^{(c)})$ and the reconstructed data $\tilde{X} \equiv (\tilde{X}^{(1)}, \tilde{X}^{(c)})$ forms the loss objective we wish to minimize. Through minimizing this difference, the decoder and encoder can *self-supervise* the learning of the dataset's latent representation Z_1 and the accuracy of the reconstructed data $\tilde{X}$.

The training is optimized by simultaneously minimizing the self-reconstruction error and the cross-reconstruction error (see Subsection 1.3.1). At its core, the learning mechanism occurs via the neural network parameterization in the encoder and decoders.

Let $p_\theta(\mathbf{z}|\mathbf{x})$ be the posterior/decoded latent parameters $\mathbf{z}$ conditional on data $\mathbf{x}$, and let $p_\theta(\mathbf{x})$ be the marginal likelihood, such that

$$\mathbf{x} \sim p_\theta(\mathbf{x}). \tag{1.1}$$

The marginal distribution, also referred to as the marginal likelihood, is

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z}, \tag{1.2}$$

where $p_\theta(\mathbf{x}, \mathbf{z})$ denotes a deep latent variable model whose prior distributions are flexibly and nonparametrically formed by normalizing flow (see Subsection 1.3.1). We optimize the variational parameters ϕ such that

$$q_\phi(\mathbf{z}|\mathbf{x}) \approx p_\theta(\mathbf{z}|\mathbf{x}). \tag{1.3}$$

The optimization is done with a loss function, which is derived from the log-likelihood of the data (Kingma and Welling, 2019):

$$\begin{aligned}
\log p_\theta(\mathbf{x}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x})] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})}\right]\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})}\right]\right] \\
&= \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right]\right]}_{=\mathcal{L}_{\theta,\phi}(\mathbf{x})\equiv\text{ELBO}} + \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})}\right]\right]}_{=D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}|\mathbf{x}))}
\end{aligned} \tag{1.4}$$

We want to maximize the log-likelihood of observing the data. Thus, we derived two terms from Equation 1.4:

- A latent loss, in the form of Kullback-Leibler (KL) divergence D_{KL} between the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and the actual posterior $p_\theta(\mathbf{z}|\mathbf{x})$. The KL Divergence is non-negative,

  $$D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x})) \geq 0, \tag{1.5}$$

  and in standard VAE, also known as “Vanilla VAE", is parameterized to be “close to” the Normal distribution $N(0, 1)$ in order to keep the divergence suitably small. However, this approximation to $N(0, 1)$ severely limits the expressiveness of the encoding, and therefore we alleviate this restriction via normalizing flows in Section 1.3.1.

- The “variational lower bound", or “evidence lower bound" (ELBO) $\mathcal{L}_{\theta,\phi}(\mathbf{x})$. Its name is derived from the fact that, due to the non-negativity of the KL Divergence, the ELBO acts as a lower bound on the log-likelihood of the data:

  $$\begin{aligned}
  \mathcal{L}_{\theta,\phi}(\mathbf{x}) &= \log p_\theta(\mathbf{x}) - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x})) \\
  &\leq \log p_\theta(\mathbf{x})
  \end{aligned} \tag{1.6}$$

Optimizing Evidence Lower Bound (ELBO).

Re-organizing equation 1.4 shows that maximizing the ELBO will optimize two measures of interest:

- Maximization of the marginal log-likelihood of $p_\theta(\mathbf{x})$;
- Minimization of the KL Divergence, therefore the encoded approximation $q_\phi(\mathbf{z}|\mathbf{x})$ becomes closer to the true posterior $p_\theta(\mathbf{z}|\mathbf{x})$.

Maximizing the ELBO ($\mathcal{L}_{\theta,\phi}(\mathbf{x})$) will therefore be the objective function with which each of the VAEs will be constructed. In practice, this is done by implementing mini-batch stochastic gradient descent optimization in which the data are split into mini-batches of random samples from the original dataset. In each step, the algorithm computes the reconstruction loss on mini-batch $B = \{x_1, ...x_N\}$ and estimates the gradient $g_B = \frac{1}{|B|}\sum_{i=1}^{N} \nabla_{\theta,\phi}\mathcal{L}_{\theta,\phi}(x_i)$. Then θ and ϕ are updated following the gradient direction $-g_B$. This will allow the model to approach the local minimum of $-\mathcal{L}_{\theta,\phi}(\mathbf{x})$, thus optimizing the VAE. It is in the SGD that differential privacy will be implemented; however, we first describe the rest of the data fusion process – improvement of VAE using normalizing flows and the fusion process with bidirectional transfer-learning.

Normalizing Flows.

One challenge of fitting VAEs is that they are limited in their ability to capture the data generating process. Specifically, VAEs perform encoding using a univariate, Normal prior, $N(0,1)$, due to the construction of the loss function (specifically, due to Kullback-Leibler (KL) divergence D_{KL}).

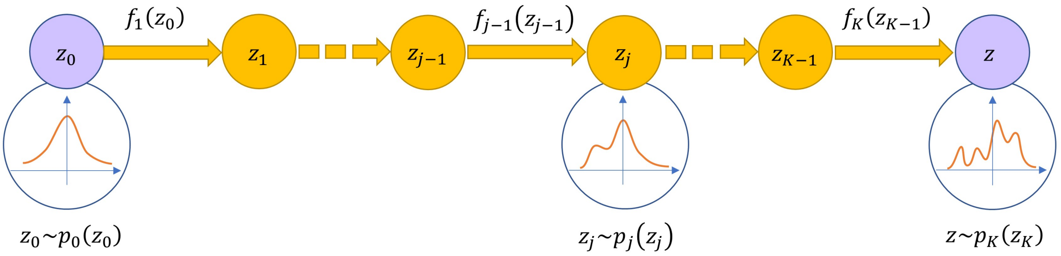


FIGURE 1.3: Illustration of normalizing flow – a series of bijective functions $\mathbf{z} = f_K \circ ... \circ f_2 \circ f_1(\mathbf{z_0})$ allows to flexibly represent the data.

To overcome this challenge, we allow the encoder of each dataset to be flexibly formed using a normalizing flow architecture. A normalizing flow (D. Rezende and Mohamed, 2015; Papamakarios et al., 2021) forms a sequence (flow) of non-linear, bijective (volume-preserving, invertible) transformations. These transformations are composed onto a draw from a simple distribution (e.g., the standard Normal distribution), making it a more accurate (yet complex) representation of the underlying data generating process. Normalizing flows have been used to improve the expressiveness and accuracy of a multitude of deep learning methods. For recent review papers, we refer the readers to Kobyzev, Prince, and Brubaker, 2020 and Papamakarios et al., 2021.

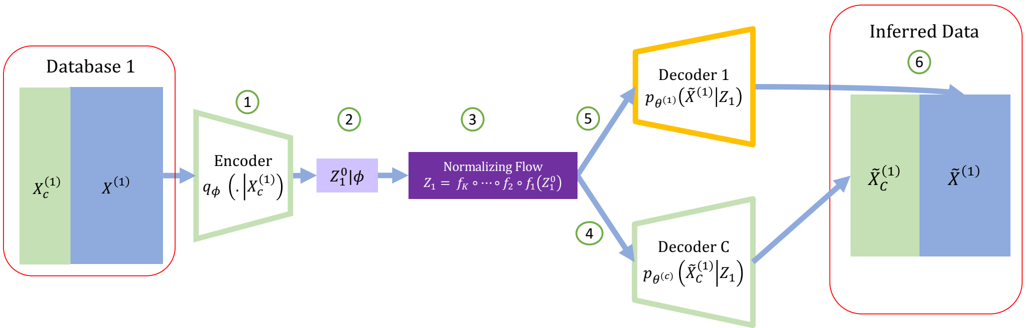

Database 1
$X_c^{(1)}$
$X^{(1)}$
1
Encoder
$q_\phi\left(\cdot\middle|X_c^{(1)}\right)$
2
$Z_1^0|\phi$
3
Normalizing Flow
$Z_1 = f_K \circ \cdots \circ f_2 \circ f_1(Z_1^0)$
5
Decoder 1
$p_{\theta^{(1)}}(\tilde{X}^{(1)}|Z_1)$
4
Decoder C
$p_{\theta^{(c)}}\left(\tilde{X}_C^{(1)}\middle|Z_1\right)$
Inferred Data
6
$\tilde{X}_C^{(1)}$
$\tilde{X}^{(1)}$

FIGURE 1.4: Illustration of a single VAE with normalizing flow. Here, the components (1,2,4,5,6) are identical to the ones in Figure 1.2. The only difference is the added normalizing flow component (3) that model allows for greater complexity in the data relationships through bijective functions.

Figure 1.3 illustrates a normalizing flow[4], whereas Figure 1.4 illustrates where the normalizing flow will be incorporated: Instead of having a simplified latent encoding, distributed $\mathbf{z} \sim N(0,1)$, we add in an intermediate series of bijective functions and a latent encoding $\mathbf{z_0} \sim N(0,1)$, such that

$$\mathbf{z} = f_K \circ ... \circ f_2 \circ f_1(\mathbf{z_0}). \tag{1.7}$$

This allows the resultant latent parameters $\mathbf{z}$ to flexibly capture data relationships of greater complexity. In turn, this enables our encoder and decoders to represent the joint distribution more accurately in the data fusion process.

Figure 1.5 illustrates the improvement of the reconstruction of a sample of digits using normalizing flow. VAEs (Vanilla VAEs in this case; Kingma and Welling, 2019) are known to create blurry images (D. J. Rezende and Viola, 2018) due to the limitation described in Subsection 1.3.1. The use of a normalizing flow (in this case $K = 7$ layers of bijective transforms) allows a richer underlying regularization distribution and results in much clearer images that capture the original digits well.

1.3.2 Learning Mechanism

In the description of the VAE, we detailed the creation of a single VAE for each dataset we fuse. This explanation omitted an important part of the process, the data fusion itself. If each dataset is encoded separately, where does the “magic" of data

[4] The illustration in Figure 1.3 is inspired by Lilian Weng: https://lilianweng.github.io/lil-log/2018/10/13/flow-based-deep-generative-models.html.

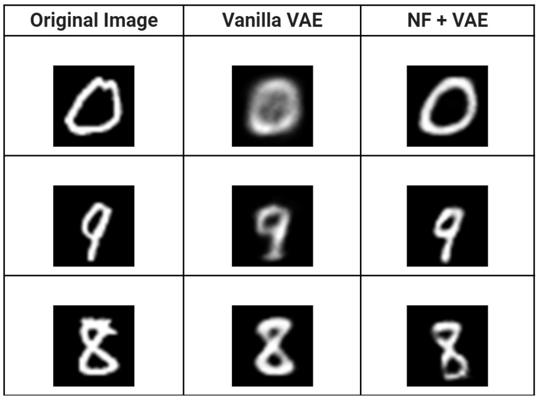


FIGURE 1.5: Results of VAEs with and without normalizing flow: results of a VAE (middle column) and a VAE with normalizing flows (VAE + NF, right column). The first VAE creates a blurry image, whereas the VAE with NF is much clearer.

fusion occur? In this subsection, we further explain the learning mechanism behind the data fusion.

To achieve the goal of learning the joint representation of the fused data while reducing privacy risks, PPDF proceeds as follows: A unique encoder takes as inputs the two datasets $X^{(1)}$ and $X^{(2)}$ including their common variables $X_c^{(1)}$ and $X_c^{(2)}$; the encoder is amortized only by the common variables $X_c^{(1)}$ and $X_c^{(2)}$. The encoder emits two sets of latent encodings, Z_1^0 and Z_2^0, that are further modified through normalizing flows (see Subsection 1.3.1). The cross-reconstruction of dataset $\tilde{X}^{(2)}$ (respectively, $\tilde{X}^{(1)}$) with respect to dataset $\tilde{X}^{(1)}$ (resp. $\tilde{X}^{(2)}$) is operated by plugging the latent enconding Z_1 (resp. Z_2) into Decoder 2 (resp. Decoder 1). A third decoder, Decoder C, reconstructs the common variables present in both datasets using a similar manipulation, but it is parameterized by a single set of parameters $\theta^{(c)}$ that learns the shared representation of the common variables across datasets. Conversely, Decoder 1 and Decoder 2 are parameterized by separate sets of parameters $\theta^{(1)}$ and $\theta^{(2)}$ that learn idiosyncratic representations of each dataset. The bottle-necking of the architecture through the one encoder and three decoders allows us to generalize the common learning space into the reconstructed full data.

Note that the encoding is merely a representation of the joint distribution, does not include the raw data, which might be identifiable, and is differentially private, as will be explained in Subsection 1.3.3. The encoding and the differential privacy mechanism assure that only differentially private latent representations of the data, and not raw data, are passed to receiving parties.

Once the encoder and decoders are optimized, a query based on any subset of variables can be made onto the joint distribution of the remaining variables across both datasets, which is the primary objective of data fusion.

Figure 1.6 illustrates the full PPDF architecture. Based on the conceptual framework of Bayesian canonical correlation analysis (BCCA), we treat each dataset as a

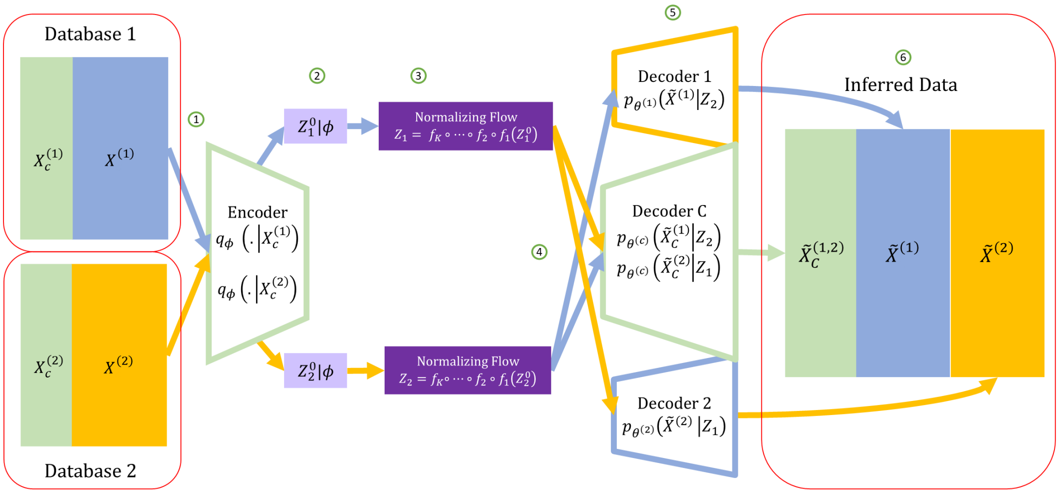

Database 1
$X_c^{(1)}$
$X^{(1)}$
Database 2
$X_c^{(2)}$
$X^{(2)}$
1
Encoder
$q_\phi\left(\cdot\middle|X_c^{(1)}\right)$
$q_\phi\left(\cdot\middle|X_c^{(2)}\right)$
2
$Z_1^0|\phi$
$Z_2^0|\phi$
3
Normalizing Flow
$Z_1 = f_K \circ \cdots \circ f_2 \circ f_1(Z_1^0)$
Normalizing Flow
$Z_2 = f_K \circ \cdots \circ f_2 \circ f_1(Z_2^0)$
4
5
Decoder 1
$p_{\theta^{(1)}}(\tilde{X}^{(1)}|Z_2)$
Decoder C
$p_{\theta^{(c)}}\left(\tilde{X}_C^{(1)}\middle|Z_2\right)$
$p_{\theta^{(c)}}\left(\tilde{X}_C^{(2)}\middle|Z_1\right)$
Decoder 2
$p_{\theta^{(2)}}(\tilde{X}^{(2)}|Z_1)$
6
Inferred Data
$\tilde{X}_C^{(1,2)}$
$\tilde{X}^{(1)}$
$\tilde{X}^{(2)}$

FIGURE 1.6: Detailed architecture of PPDF – two VAEs and each with its own normalizing flows. A single encoder (1) is amortized only by the common variables $X_c^{(1)}$ and $X_c^{(2)}$. The encoder emits two sets of latent encodings, Z_1^0 and Z_2^0 (2), that are further modified through bijective transformations (3), i.e. via normalizing flows. The cross-reconstruction of dataset $\tilde{X}^{(2)}$ (respectively, $\tilde{X}^{(1)}$) with respect to dataset $\tilde{X}^{(1)}$ (resp. $\tilde{X}^{(2)}$) is operated by plugging the latent encodings Z_1 (resp. Z_2) into Decoder 2 (resp. Decoder 1). A third decoder, Decoder C, reconstructs the common variables present in both datasets using a similar manipulation (4), but is parameterized by a single set of parameters $\theta^{(c)}$ that learns the shared representation of the common variables across datasets. Conversely, Decoder 1 and Decoder 2 are parameterized by separate sets of parameters $\theta^{(1)}$ and $\theta^{(2)}$ that learn idiosyncratic representations of each dataset. The bottle-necking of the architecture through the one encoder and three decoders allows the induction of a common learning space. We note that this architecture allows for differently sized datasets 1 and 2.

multivariate random variable with unknown parameters. The encoder – with the flexibility of normalizing flows – encodes each dataset into a latent representation. The encoding task is performed using the common variables only, but the representation is of both the common and unique variables of each dataset. The common variables allow us to construct the mapping between the two datasets in a formation of a joint latent representation.

Using the augmented variational parameters, data fusion (i.e., cross - reconstruction) can then occur as probabilistic imputations from a joint posterior predictive distribution. In other words, by garnering the marginal posteriors $p_{\theta^{(1\text{ or C})}}(X^{(1)}, X_c^{(1)}|Z_1)$ and $p_{\theta^{(2\text{ or C})}}(X^{(2)}, X_c^{(2)}|Z_2)$ – where Z are the common latent representations, $X^{(m)}$ are the dataset-specific variables, and $X_c^{(m)}$ are the shared variables – we can obtain the posterior predictive joint distribution $f(X^{(1)}, X^{(2)}, X_c^{(1)}, X_c^{(2)})$.

The end result of this data fusion is that, for every entry in either dataset, we have a probabilistic reconstruction of the matching entry from the other, such that with $p({\tilde{X}_i}^{(1)}|X_i^{(2)})$ we can construct $[\tilde{X}_{i,c}^{(1)}, X_{i,c}^{(2)}, \tilde{X}_i^{(1)}, X_i^{(2)}]$, and conversely, inferring Dataset 2 onto Dataset 1, we have $p(\tilde{X}_i^{(2)}|X_i^{(1)})$ begetting $[X_{i,c}^{(1)}, \tilde{X}_{i,c}^{(2)}, X_i^{(1)}, \tilde{X}_i^{(2)}]$.

As another conceptual illustration, our VAE can be thought of as the writing of a *Rosetta stone* by one linguist and three scribes: the reader can think about the two datasets as two foreign languages, where the common variables act as a common syntax, vocabulary, and grammar, but without an exact translation across the two datasets, which would require sharing the same rows across the two datasets. The two languages are initially unfamiliar to the linguist and the three scribes, but the linguist's job (encoder) is to study these two languages and their common concepts as inputs and encode a shared representation (books) as latent variables Z_1 and Z_2 through the common variables, parameterizing this representation with ϕ. This process allows the encoding of higher levels of abstraction between the two languages, as if the linguist was writing her interpretation of the two languages to be used by the scribes. Normalizing flows help to reach the adequate level of abstraction. Then, the scribes' task (the three decoders) is to go back to a lower level of abstraction and respectively reconstruct the original two languages and their common elements using the linguist's books (the latent encodings). The three scribes decode their own learning from the linguist's books using parameters $\theta^{(1)}, \theta^{(2)}, \theta^{(c)}$ for the two original languages and their common elements, respectively. The linguist knows that her books will be used by the scribes, so she tries to convey as much information about the other language in each book. Once the model is trained, it is as if we wrote a Rosetta stone that allows the analyst to go from one language to the other using the bottle-necking via the unique encoder and the three decoders.

Using the combination of VAE, NF and the mutual learning mechanism, our framework is capable of fusing two datasets – or more – learned as a single joint distribution. However, as of yet, we have not described the privacy enhancing methodology. In the next subsection, we introduce the differential privacy component within the process.

1.3.3 Privacy Preservation Measures and Controls

One of the essential parts of the proposed methodology is the ability to preserve users' privacy. There is an inherent tension between privacy and accuracy, when it comes to fusing datasets. The best data fusion will match each user's variables in a dataset to the same user's variables in the other dataset. However, such matching might reveal users' full sets of attributes, and, in some cases, will allow the researcher to uniquely identify them along with traits they did not choose to disclose or that can potentially harm them. This risk is known as a "linkage attack," and has been demonstrated by Sweeney, 1997 and Narayanan and Shmatikov, 2008, among others.

On the other side of the privacy-accuracy trade-off, a completely private data fusion might include only the summary statistics of the variables of the entire population in one dataset and merely correlate them with those of the other dataset. This will allow for learning of the joint distribution for the entire population but will not allow for heterogeneity and covariance across datasets.

Consider the telecom carrier's CRM dataset, along with a detailed, anonymous survey on attitudes and past behaviors conducted by an external company. Any identification that results from data fusion might harm individuals' expectations of privacy: they may have wished to stay anonymous, not revealing their individual attitudes or behaviors. Moreover, intellectual property of the external survey may demand the anonymity of the respondents. Therefore, in such cases, the data holders (the telecom carrier in this case) may prioritize privacy over accuracy.

As another illustrative example, less privacy might be deemed necessary when handling datasets from public sources, since it is reasonable to assume that, by the mere presence of their data in a public dataset, individuals are not expecting privacy guarantees[5].

Therefore, it is up to the data holders to assure they are in line with customers' expectations, intellectual property considerations, regulations, privacy policies, and known risks when using a data fusion method.

As part of the proposed privacy preserving methodology, we offer tuning mechanism that will enable the data holder(s) a higher sense of control over the level of privacy vs. accuracy. These privacy guarantees, with tuning mechanism, are achieved using differential privacy.

Differential privacy, first introduced by Dwork, McSherry, et al., 2006, is a data privacy mechanism used to formalize the trade-off between privacy and accuracy through the introduction of added noise during model training. It allows the researcher to tune the risk associated with identifying a person from a dataset, and explicitly set a "privacy budget." Differential privacy is considered state-of-the-art among current privacy preserving methodologies and has been used for data publication or data release (Takagi et al., 2020; Fung et al., 2010; Narra and Chiyuan Zhang, 2022), including in the release of data from the U.S. 2020 Census (see explanation from U.S. Census Bureau 2021).

Differential privacy relies on the assumption that if it is impossible to ascertain that any particular user's data were used in an analysis, then their privacy is preserved. From another angle, a model-based analysis is considered differentially private if each individual has a bounded probability to be determined as included in the analysis's training dataset, relative to another dataset that only differs in the removal of their data. Differential privacy therefore relates to the assurance (up to a bounded probability), that the inclusion of an individual in a dataset will not change

[5] While it is reasonable to assume that privacy expectations are low, the researcher might still want to err on the side of caution and choose to de-identify individuals.

the marginal outcomes of a model-based analysis (e.g., coefficient estimates, standard errors, posterior predictives) relative to a dataset that does not include their data.

There are two realizations of differential privacy, ε and δ; we first begin by defining ε-differential privacy. Consider two adjacent datasets, D and D', that are the same except that dataset D' has one more observation, i.e., $D' = D \cup x_i$ where x_i are the data of individual i.

An algorithm M is considered ε-differentially private ($\varepsilon \in \mathbb{R}$ and small as desired), if for every output S, we receive the same output S with the dataset D' at a probability that is at most e^{ε} that of dataset D. A low ε means that for the two datasets that differ only in the existence of x_i's data, we have very low probability of distinguishing between the two outputs. This makes inclusion of x_i in the data very hard to detect:

$$\Pr\left(M(D) \in S\right) \leq e^{\varepsilon} \cdot \Pr\left(M(D') \in S\right). \tag{1.8}$$

This can also be seen as: survey respondent i cannot be revealed as an input to our model, if they haven't responded to it. The probability of being identified as a respondent, through a variation in the outputs of algorithm M, would be very low in such a case. ε-differential privacy would allow us to state that even if i **is** a respondent to the survey, the probability of their being identified as a data input is very low as well; it is at most e^{ε} more likely. ε is therefore a measure of the privacy loss that the marginal impact of a single customer's data onto a model can be uniquely identified back to that customer. By construction, smaller values of ε would lead to lower privacy loss (i.e., higher privacy guarantees).

As another variation of differential privacy, Dwork, Kenthapadi, et al., 2006 added an upper bound of the individual risk δ, such that:

$$\Pr\left(M(D) \in S\right) \leq e^{\varepsilon} \cdot \Pr\left(M(D') \in S\right) + \delta. \tag{1.9}$$

The addition of $\delta \in \mathbb{R}_{\geq 0}$ serves as a "failure probability", which acts as a tolerance to the risk associated with identification – allowing for the possibility that ε-differential privacy is broken with probability δ. Intuitively, such failures can arise as a result of a data breach, whereby the dataset, partially or in its entirely, is exposed to unauthorized parties. As such, common privacy budgets are set to have $\delta < \frac{1}{|D|}$, where $|D|$ is the number of users in the dataset. Thus, in addition to the likelihood of revealing an individual's identify from model outcomes, such exposure may also arise from factors unrelated to M, such as data leaks.

In line with state-of-art work in this area, we implement differentially private doubly stochastic variational inference algorithm for training, in its (ε, δ) form (Jälkö, Dikmen, and Honkela, 2017), which is derived from DP-SGD (Abadi et al., 2016).

Practically speaking, we incorporate the *d*3*p* package for differentially-private probabilistic programming (Prediger et al., 2022), which provides a reliable, high - performance implementation of the algorithm. This open source software is usually used for data publication and in our case is extended to data fusion.

At each step of the PPDF training, we compute the gradient of the loss function $g(\mathbf{x}) = \nabla_{\theta,\phi}\mathcal{L}_{\theta,\phi}(\mathbf{x})$, or, for a random subset of samples $B = \{x_1, ..., x_N\}$, compute the gradient of this mini-batch: $g_B = g_t(x_i) = \frac{1}{|B|}\sum_{i=1}^{N}\nabla_{\theta,\phi}\mathcal{L}_{\theta,\phi}(x_i)$. The parameters are then updated following the gradient with learning rate η_t, such that the updating of parameters θ, ϕ is $\{\theta, \phi\}_{t+1} = \{\theta, \phi\}_t - \eta_t \cdot g_t$. This is a procedure common for estimating VAEs, but DP-SGD adds two more steps in the computation of the gradient, to assure privacy.

1. Clipping the norm of each gradient g_t, to assure that the information of each individual in a mini-batch is limited:

$$\bar{g}_t(x_i) = \frac{g_t(x_i)}{\max\left(1, \frac{\|g_t(x_i)\|_2}{C}\right)} \tag{1.10}$$

2. Adding noise from a Normal distribution such that

$$\tilde{g} = \frac{1}{|B|}\left(\sum_i \bar{g}_t(x_i) + \mathcal{N}\left(0, \sigma^2 C^2 \mathbb{I}\right)\right). \tag{1.11}$$

The parameters for the clipping of the norm and for the added noise are computed based on the desired ε and δ, in a process referred to as "privacy accounting", detailed by Abadi et al., 2016. As noted, δ is typically set to a value less than $\frac{1}{|D|}$ (i.e., the tolerance for unexpected failures are inverse to the size of the dataset), and from this point, lower ε (i.e., a more stringent privacy budget) results from larger norm clipping and additive noises.

Now having explicated the mathematical foundations of the three core components of PPDF – variational autoencoders, normalizing flows, and differential privacy – we next turn our attention to how PPDF, as a data fusion technique, handles missing data and sample selection biases across datasets.

1.3.4 Handling Missing Data and Selection Bias

Inherently, every data fusion task is intended to impute missing values; the researcher is imputing the unique variables from one dataset into the other, relying on the common variables in both. There can be several types of missingness that can confound any exercise in the combination of datasets, including data fusion. Rubin, 1976 classifies three mechanisms of missing data: missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR, also known as non-ignorable). For an expanded discussion, see Appendix A.1.

In almost all realistic datasets, there is a need to overcome the problem of missing values within variables. Missingness is a common problem in the social sciences and, in particular, in marketing research. Collecting data from human subjects is highly likely to result in missing information. This occurs for a variety of reasons: unwillingness of users to respond to some questions (Bradburn et al., 1979); changes in experimental design over time, which might also result in missing observations for whole variables (Graham, 2009), and flaws in data collection carried out in certain field settings.

In line with state-of-art data fusion methods (Gilula, McCulloch, and Rossi, 2006; Qian and Xie, 2022), PPDF handles missingness of types MAR or MCAR via imputation (referred to as "cross-imputation" in our framework). Moreover, any cross-imputation undertaken by PPDF across datasets is equivalent to a sampling-adjusting correction (Elea M Feit and Bradlow, 2021) on data that are MNAR, assuming that any variable that affects the non-representative missingness is observed amongst the datasets (i.e., selection on observables). Specifically, PPDF nonparametrically completes (augments) a latent representation of both within-variable missingness and the obvious whole variable missingness, with observed and inferred variables from the other dataset.

As we will show in Section 1.5.3, in the case of the telecom carrier's survey and CRM data fusion, customers with more extreme attitudes (in terms of very high or very low Likelihood to Recommend) were also more likely to respond to the survey (this pattern can also be found in online reviews; see e.g., Schoenmueller, Netzer, and Stahl, 2020). Such self-selection leads to an empirical distribution on survey outcomes that is not necessarily representative (i.e., biased) of those of the full customer base. This is because we are *missing* responses proportionate to the customers who have less extreme attitudes and are therefore less likely to respond. In order to obviate this selection bias in the survey, PPDF will nonparameterically adjust empirical distribution of the survey outcomes to reflect that of the full customer base via cross-imputation of expected survey responses. That is, PPDF infers individual-level survey responses for the full customer base, even if these customers have not responded to the survey.

1.4 Simulation Exercises

Before moving onto our focal managerial applications, in this section, we showcase the ability of PPDF to fuse datasets for which we know the underlying joint distribution, with varying tuning parameters and differential privacy specifications.

1.4.1 Sensitivity Analysis 1 – MNIST Digits Data

The first simulation is based on MNIST data. Described by Deng, 2012, the MNIST dataset is frequently used to assess classification methods. It includes 60K black and white images of numeric digits, each with $28 \times 28 = 784$ pixels. In Figure 1.5, we demonstrated the improvement of VAE+NF over a standard VAE in the reconstruction of the digits.

To illustrate PPDF for the purpose data fusion, in the following simulation, we split each image into two – allocating a portion of the middle pixels as if they were common variables – and will then fuse them back. Specifically, we left 300 pixels from the center of each digit to be the common variables for each observation, and $\frac{784-300}{2} = 242$ pixels were considered unique variables for each dataset[6].

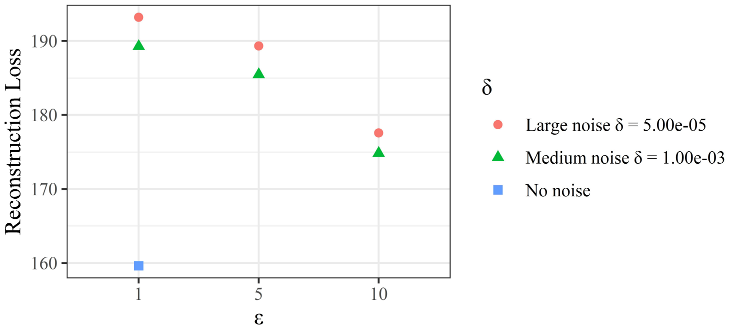


FIGURE 1.7: Reconstruction loss (the loss relative to the original images) as a function of δ and ε. Larger noise (smaller δ and smaller ε) represents lower tolerance to re-identification. No noise means no privacy guarantees at all and acts as a reference. The learning rate is $\eta = 2 \cdot 10^{-4}$ and common variables are 300 of the 784 pixels in each digit.

In order to highlight the roles of δ- and ε- differential privacy in varying the noise levels of the resultant fusion, Figure 1.7 shows the reconstruction loss for varying levels of noise.

The smallest added noise was no noise at all (blue square), which acts as a reference. As we vary ε from larger to smaller values, we limit the privacy budget, increasing privacy by adding more noise. This results in higher levels of reconstruction loss. Similarly, as we vary δ, we add more noise, which results in higher loss.

Following the fusion, we get the reconstructed images presented in Table 1.1: The upper row corresponds to the basic data fusion model with no added noise, and the rest of the rows show the resultant images with the varying levels of ε and δ.

As in any data fusion, the ability to reconstruct missing data depends on the number of common variables. This is usually a given, but if a company wishes to split a dataset in order to protect its customers, it can potentially control the number of common variables before splitting. Figure A.2 in Appendix A.2 shows the ability to reconstruct images with a varying number of common variables (pixels in MNIST images). The more commonality there is between datasets, the better the reconstruction and the faster the convergence rate. Data holders who wish to split datasets may try their split (vary not only the number of common variables, but also which

[6] We varied the number of common digits later, see Figure A.2.

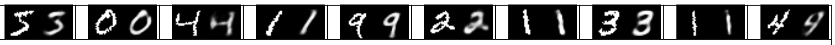
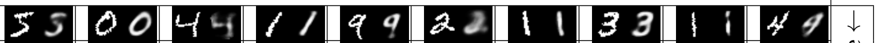
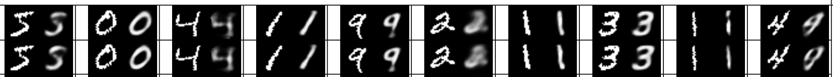
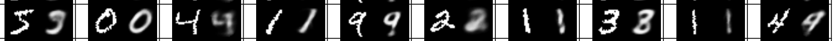
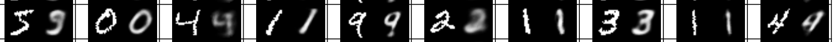
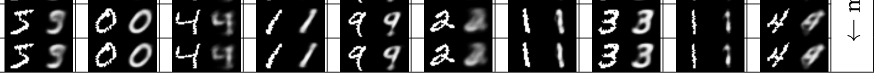
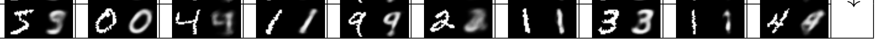
No noise			
ε	δ		
10	1^{-3}		← more noise ←
10	5^{-5}		
5	1^{-3}		
5	5^{-5}		
1	1^{-3}		
1	5^{-5}		

TABLE 1.1: Results of data fusion – MNIST dataset: An example of 10 MNIST digits with varying levels of noise added. The left side of each pair of digits is the original digit. The right side is the reconstructed digit after splitting and fusing the pixels: the middle 300 pixels are common across datasets, and the rest are unique to each of the two datasets. The upper row has no added noise. The next rows have varying levels of ε and δ. A smaller ε represents a lower privacy budget; therefore, more noise is added to the DP-SGD, as explained in Subsection 1.3.3. Conceptually, for hand-written digits, more privacy means a decreased ability to recognize the specific instance of the digit by removing the identifiers of the it.

variables remain common) and test the reconstruction loss with varying ε and δ to determine the privacy measures in different contexts.

Beyond the tuning of the DP parameters, other parameters can be tuned to improve reconstruction loss. Some of these relate to the underlying structure of the VAEs – namely the size of the vector Z of latent encoding or the size of the hidden layer in the VAE's inference and amortization networks. This are displayed in Figure A.2 in Appendix A.2.

1.4.2 Sensitivity Analysis 2 – Survey Data

The second simulation, which showcases sensitivity to different parameter specifications, is based on data from a survey with 5.5K participants[7]. The survey included 61 questions. We split the data column-wise, into two datasets, X_1 and X_2. There were 28 common variables in this analysis.

We first conducted a sensitivity analysis to the tuning parameters without differential privacy. Figure A.3 shows the mean absolute error relative to the original data, across all columns of the joint dataset, after running for 10K epochs, and for various tuning parameters.

Following the selection of tuning parameters unrelated to differential privacy, we now illustrate the sensitivity of PPDF to the addition of privacy preserving guarantees. For this, we used the best parameters found in the No-DP analyses and added noise in varying scales. Figure 1.8 presents the mean absolute error of runs with varying ε values. Figure 1.8 also shows the benchmark of the No-DP model for

[7] This survey is taken from a customer base unrelated to the the telecom carrier data used for our main application.

comparison. From this figure it is apparent that as the noise gets bigger, accuracy is reduced, as expected.

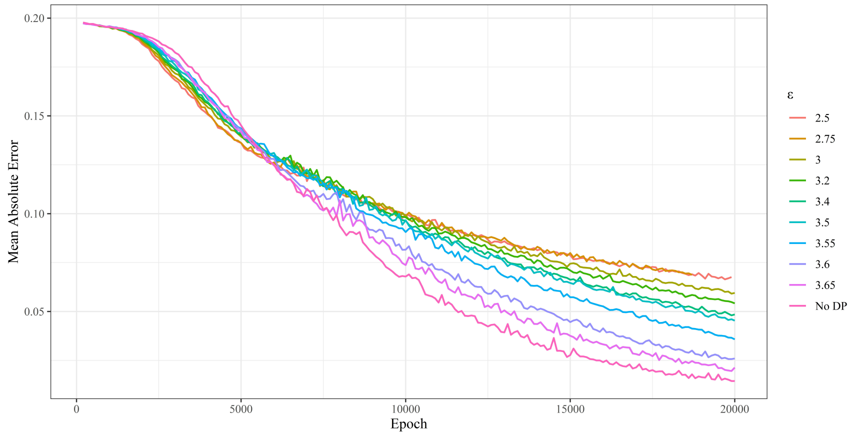


FIGURE 1.8: Mean absolute error of PPDF on simulated data, under multiple specifications of ε. In this plot, the other tuning parameters were chosen as they performed best in the aforementioned (non DP) analyses: batch size = 256, learning rate = 1e-4, Z dimension = 50, and $\delta = 1e - 4$ when applicable.

A note on computational intensity based on the described sensitivity analyses: One concern in estimating large learning models such as ours is the level of scalability and length of run-times. To ease this concern, we highlight that each of the sensitivity analyses on the 5.5K survey responses took at most 10 minutes for 20K epochs, on a commercial GPU (Google Colab), widely available for researchers and managers alike. The analyses on MNIST digits were conducted on a personal computer with 16G RAM and took 23 minutes to complete. While this is a perfectly reasonable run-time, progress is constantly being made on processing units, and we expect an even further reduction in these run-times in the near future.

1.5 Application: Telecom Carrier Anonymous Survey and CRM Data

In the previous section, we illustrated the ability of PPDF to fuse MNIST digits and performed a synthetic simulation of survey re-fusing under varying privacy budget constraints. We now present an application of PPDF and conduct a data fusion of two datasets provided by a leading U.S. telecom carrier.

1. Detailed sample from CRM data randomly selected from the telecom carrier's full customer base (3.6 million accounts, which are roughly 2.0% - 5.0% of the full customer base, excluding pre-paid accounts).[8]

[8] Confidentiality: the data were provided to the researchers under a strict non-disclosure agreement. Some figures and identifiers of the telecom carrier and of their customers are removed or obscured. The percentage range above is with respect to total customers across all three major U.S. telecom carriers at the time of writing.

2. The responses from an industry-standard, anonymous customer satisfaction survey, hereafter, the “external survey". The survey was conducted by an external surveying company (hereafter the “surveying company") that continuously conducts market research on customer satisfaction, on behalf of the market leaders in the wireless telecom sector. In particular, survey outcomes include measures on satisfaction with different service components, as well as an overall Likelihood to Recommend[9]. We provide a set of questions in the likeness of those in the external survey in Appendix A.3.

The external survey and the resulting measures are well-known industry standards and are being used by customers and managers to assess quality of service. It is therefore of utmost importance for the firm and their competitors in the telecom space to improve this customer satisfaction score to both improve their customer well-being and to attract new customers. Our goal in this data fusion exercise is to combine the CRM data with the survey data to gather insights on the relationship between stated satisfaction and actual engagement with the company.

Novel to this research, for the purpose of assurance of proper fusion, an additional internal calibration survey was conducted by the telecom carrier. This survey is identical to the “external survey” conducted by the external surveying company, except that it was conducted by the firm itself. The email link was unique to each recipient, and so user IDs were collected through the associated link. This internal calibration survey provides a critical, one-of-a-kind ground-truth mechanism to ensure our methodology properly fuses unique dataset by allowing for retrospective testing of fusion accuracy when user identities are definitively known and overlap across datasets. However, of note, this step is not necessary for future end-users of PPDF and is presented here solely to convey the proposed methodology’s accuracy.

The telecom carrier’s stated goal in this process is to explore the relationship between customer satisfaction and the likelihood to churn from the company’s services. Customer surveys, while extensive and include many engagement variables, do not include measures of financial outcomes, technical metrics on device and network usages, nor plan and service details. For the firm’s managers, the expected outcome of dataset fusion (of the external survey outcomes with the extensive CRM data they hold. The survey allows the detection of pitfalls in customer satisfaction as defined in the survey and further understood based on the respective behaviors and detailed engagement represented in the CRM data – behaviors that may not be well represented in the survey questions due to its length, false memory of the respondents, and selection to report or not to report certain experiences, among other reasons.

[9] LTR is a ubiquitous survey metric, from which Net Promoter Score (NPS) is often calculated. NPS, while utilized at the firm, is precluded from our analysis, where we focus directly on LTR outcomes.

Crucially, the external survey data should remain anonymous and any identification of customers as respondents of the survey, must be avoided to ensure continued customer trust and voluntary participation in the surveying process in the future. Therefore, to avoid the risk of identification, we use PPDF to fuse the data sets while preserving anonymity.

We begin by describing the survey, followed by a description the CRM data. We then show the outcomes of PPDF with and without normalizing flows, and with and without differential privacy. Finally, we show some insights elucidated from the joint distribution of the fused data.

1.5.1 Internal and External Survey Protocol and Data

Both internal and external surveys were conducted in the third quarter of 2021. For the internal survey, an email was sent to 2.1 million of the telecom carrier's customers (randomly selected from the entire customer base, except for pre-paid customers that were excluded). Emails were sent gradually throughout the surveying period. Approximately 20K customers responded to the survey, corresponding to a response rate of 0.95%. Participants did not receive compensation for completing the survey.

The external survey was conducted by an external surveying company that continuously conducts market research on customer satisfaction on behalf of the market leaders in the wireless telecom sector. Approximately 8K respondents responded to the external survey, all of whom were customers of the telecom carrier at the time of responding.

Both surveys were identical and included customer satisfaction/perception questions, questions about engagement with the firm's products and services, questions about the customer relationship with the telecom carrier, and socio-demographic questions. We screened out all who started but did not complete the survey but there were few of those.

In conjunction with the telecom carrier, we delineated the survey taxonomy into four mutually exclusive question types:

1. Identifiers (e.g., socio-demographics)

2. Relationships (e.g., plan choice, account type, devices)

3. Engagements (e.g., times a retail store was visited, plan-switching, customer service calls)

4. Perceptions (e.g., satisfaction with services and devices, Likelihood to Recommend (LTR))

1.5.2 CRM Data

The telecom carrier has extensive and detailed user data from the moment a user joins. The data available for this project are one of two types: common variables and uncommon variables. Common variables are shared between the survey data and the CRM data, such that a common variable for a given customer will have the same value or attribute across both datasets. Uncommon variables are idiosyncratic to their respective datasets, that is an uncommon variable exists in one dataset but not in the other. Uncommon variables may include variables related to detailed engagement with the company, services used, changes to accounts, churning, and payments. We list the common and uncommon variables in Appendix A.3.

1.5.3 Results

We first use the internal survey (20K responses), where we can deterministically link user ID to internal CRM records, to assess the trade-off between accuracy and privacy of the proposed PPDF method with and without differential privacy and compare PPDF against extant benchmark models. Next we show the outcomes of the data fusion of the external survey (8K responses) with a random sample of the CRM data (3.6 million users). Here, we focus on two customer-level outcome variables, the “Likelihood to Recommend" (LTR) response from the survey and the 12-month forward-looking churn outcome (binary) from the CRM. We highlight two advantages of PPDF: (1) a regression (i.e., discriminative) analysis that is undertaken *ex post* the generative data fusion, as an example of how PPDF enables the decoupling of data fusion and downstream analyses while still retaining full posterior predictive inference, and (2) this shows that the fused data exhibit better goodness-of-fit and explanatory power on the drivers of LTR vis-a-vis only using features from any one dataset alone. Finally, we provide managerial applications of the data fusion by relating the fused LTR scores and churn outcomes of the customer base, realized as an individual-level 'scoring' of customers with their expected LTR, a task otherwise impossible without data fusion. We also infer differences between the self-selected survey respondents vs. the customer base without self selection. On the one hand, from a substantive point of view, these results quantify the relationship between LTR and potential churn and enable managers to use LTR as a potential segmentation strategy for proactive churn management. On the other hand, from a technical point of view, these results also highlight the need to correct for selection bias in understanding customer perceptions of the company’s services and other measures of engagement.

	With Differential Privacy				Without Differential Privacy			
	CRM		Survey		CRM		Survey	
Model	Recon	Cross	Recon	Cross	Recon	Cross	Recon	Cross
PPDF	0.0532	0.0532	0.1325	0.1325	0.0380	0.0382	0.0912	0.0919
PPDF w/o NF	0.0683	0.0683	0.1422	0.1423	0.0421	0.0421	0.1028	0.1029
BCCA	0.0689	0.0689	0.1384	0.1391	0.0422	0.0422	0.1028	0.1028
SUR NN	0.0544	0.0546	0.1343	0.1347	0.0407	0.0409	0.0996	0.0998

TABLE 1.2: Goodness-of-fit (measured by mean absolute error – MAE) of variations of the proposed data fusion (with and without normalizing flows and with and without differential privacy) and two other benchmark models – Bayesian canonical correlation analysis (“BCCA") and a forward-feed neural network model (“SUR NN"). The results are split into self-reconstruction (“Recon"), namely encoding and decoding a dataset onto itself, and cross-imputation (“Cross"), which is data fusion, namely decoding of one dataset into another. Data fusion with normalizing flows provides the best results, and adding privacy-preserving guarantees increases MAE only slightly.

Model Comparison.

In Table 1.2, we provide the goodness-of-fit of the test sample, between the proposed PPDF model, with and without differential privacy,[10] against three benchmark models. These benchmarks include (a) PPDF without normalizing flows, to highlight the incremental gains from NF, (b) Bayesian canonical correlation analysis (“BCCA") (Klami, Virtanen, and Kaski, 2013), which can be understood as a nested linear-form of VAE, and (c) a forward-feed neural network model predicting survey and CRM variates from the common variables, which can be understood as a system of non-linear seemingly unrelated regressions (“SUR NN"). All models are fit on the internal calibration survey. This uniquely provides us the ground truth, one-to-one linkage between survey responses and CRM records, as discussed above. As all variables between the survey and CRM datasets are categorical in nature and are encoded in a one-hot approach,[11] we chose to report mean absolute error (MAE) across the model comparisons. In this context, MAE intuits the goodness-of-fit in terms of both absolute and percentage deviations in the test data (75%/25% random split between training and test data).

For example, in the top-left cell of Table 1.2, the MAE of the self-construction of the CRM data, is 0.0532 for the full PPDF model (i.e., trained with normalizing flows and differential privacy). This is equivalent to an out-of-sample error of 5.32% on average across all one-hot-encoded format CRM variables. Under differential privacy, the self-reconstruction error for the full PPDF model (0.1325) is more favorable than those from BCCA (0.1384) and SUR NN (0.1343). While on the other hand, PPDF without NF (0.1422) falters behind than the benchmark models, suggesting that NF

[10] In our empirical applications, we fix differential privacy to a stringent (ε, δ) budget of $(0.1, 5.0^{-5})$, which is on par with leading industry standards.

[11] For example, LTR is an eleven-point discrete scale from 0 to 10, which is then encoded as eleven binary variables. Note that dropping a category to avoid multicollinearity, while typical for regression-type model, is unnecessary in our generative modeling approach.

is critical to the VAE architecture underpinning PPDF in order to offset the loss in accuracy due to DP. While the full model without differential privacy is observed to have the best goodness-of-fit (0.0380 for CRM and 0.0912 for survey), the addition of differential privacy comes at the cost on accuracy. Taken together, data fusion under privacy is capable of generating meaningful and actionable inferences on the joint data generating process but exhibits a trade-off on accuracy compared to conventional data fusion. As we will show in the next section, despite the loss of accuracy under differential privacy, there is a tangible information gain when using the fused data in predictive exercises, while of course, enabling the privacy guarantees of DP.

Next, we have defined self-reconstruction as the ability of models to encode and decode a dataset onto itself. On the other hand, cross-imputation is the ability to decode a dataset into the other. In addition to MAEs on self-reconstruction, we show MAEs on cross-imputation (“Cross") as well (Table 1.2). Cross-imputation is the key data fusion technique used in the managerial exercises in the following sections. Namely, cross-imputation allows the telecom carrier to impute the posterior predictive distributions in survey perception outcomes (e.g., LTR) from the CRM data alone, and vice versa, based on common variables. While cross-imputation systematically results only in higher or equal MAEs across all models with or without DP, any of the observed degradations in fit are well within managerial and inferential tolerance. It is important to note that this degradation of fit is expected by the nature of generating new data. For example, for the CRM data, PPDF saw no change in MAE between self-reconstruction (0.0532) and cross-imputation (0.0532) under DP, while the increase in MAE without privacy guarantees is 0.0382 - 0.0380 = 0.0002, or 0.02%.

Having established the superior goodness-of-fit of PPDF compared to extant models, as well the accuracy-privacy trade-off of differential privacy, we next turn to applying the generative data fusion outcomes to two discriminative regressions, on LTR and churn, to illustrate the explanatory and predictive value that is reached by fusing the anonymous external survey and internal CRM data, beyond analyses on either dataset alone.

Likelihood to Recommend (LTR).

Modern customer satisfaction surveys are commonly found to be organized around the question of LTR (Keiningham et al., 2007; F. Reichheld, 2011). From the perspective of inference and analysis, the conventional usage of a satisfaction survey is to view the LTR-type perception questions as dependent variables, while other elements of the questionnaire (i.e., identifiers, relationship, engagement and usage elicitations) are the independent variables. However, as is in the case with our partner telecom carrier, most companies are keen not only to regress LTR onto these elements within the survey questionnaire, but also to connect LTR with CRM variables. For the telecom carrier, understanding the survey and CRM predictors of LTR is part of the objective of understanding their brand perception by customers as well as to rank-order existing customers by LTR.

We highlight the applicability of our framework, which is generative in nature, to aid in a downstream discriminative analysis of survey variables. This discriminative analysis elicits a great deal of interest from managers, who are particularly eager to measure and use LTR. This showcases PPDF's approach to data fusion based on VAEs: first estimating the joint data using VAEs, and only then conduct discriminative analysis. Marketing analysts are then emancipated from needing to craft complex models that must simultaneously consider data fusion and a discriminative likelihood.

Data Type	**Data Source**	**Original Data**		**DF w/o DP**		**PPDF**	
		RMSE	R^2	RMSE	R^2	RMSE	R^2
Original	Survey	3.250	0.179	–	–	–	–
	CRM	–	–	–	–	–	–
	Combined	–	–	–	–	–	–
Self-Reconstruction	Survey	–	–	3.250	0.179	3.268	0.152
	CRM	–	–	3.143	0.308	3.156	0.297
	Combined	–	–	3.100	0.346	3.127	0.324
Cross-Construction	Survey	–	–	3.255	0.171	3.266	0.155
	CRM	–	–	3.157	0.295	3.159	0.295
	Combined	–	–	3.140	0.311	3.118	0.333

TABLE 1.3: Goodness-of-fit of external survey and full CRM data

Data Type	**Data Source**	**Original Data**		**DF w/o DP**		**PPDF**	
		RMSE	R^2	RMSE	R^2	RMSE	R^2
Original	Survey	3.093	0.171	–	–	–	–
	CRM	2.887	0.393	–	–	–	–
	Combined	**2.863**	**0.411**	–	–	–	–
Self-Reconstruction	Survey	–	–	3.096	0.167	3.103	0.152
	CRM	–	–	2.899	0.384	2.988	0.307
	Combined	–	–	**2.865**	**0.409**	**2.952**	**0.340**
Cross-Construction	Survey	–	–	3.099	0.160	3.100	0.158
	CRM	–	–	2.948	0.340	2.991	0.304
	Combined	–	–	**2.873**	**0.403**	**2.953**	**0.339**

TABLE 1.4: Goodness-of-fit of internal calibration survey and surveyed CRM data

Broadly, we find that indeed incorporating CRM covariates, in addition to survey covariates, results in better explanatory and predictive power of a regression on LTR, in terms of two common discriminative goodness-of-fit metrics, root mean squared error (RMSE, where lower is better) and R^2 (where higher is better). We prefer the use of RMSE and R^2 over MAE here because they offer more stringent measures of model performance and are more sensitive to outliers. We compare RMSE and R^2 for the prediction of LTR when fusing survey and CRM data, for both the external survey (Table 1.3) and the internal calibration survey (Table 1.4). In both figures, the top-left quadrant provides these fit metrics based on the original data – that is, when LTR is regressed onto just the survey data of the full sample, just the CRM data of the full sample, and finally, both. Once again highlighting the value of the internal calibration survey, note that it is only with this survey (Table 1.4) that we can provide fit metrics for the true "original" CRM and combined datasets whereas the external survey is anonymous, and we cannot compute these metrics for the survey data. In Table 1.4, we find that RMSE is the lowest and R^2 is the highest when the LTR is regressed against the combined data.

Having established the superior goodness-of-fit of the combined data, over-and-above the individual datasets, the key question becomes – can we repeat this pattern using the decoded self-reconstruction and cross-imputation datasets? This question is pertinent in understanding whether the output of our proposed model, which is ultimately generated and synthetic in nature, can indeed capture the complexity of the underlying data generating process that begets the "sum is greater than the parts" phenomenon. Here, we once again find the phenomenon to be true for both self-reconstructed and cross-imputed data, with and without differential privacy. For the internal calibration survey (Table 1.4), combined/fused data have RMSEs of 2.865 (self-reconstruction) and 2.873 (cross-imputation), comparing favorably to the RMSEs of the synthetic survey- and CRM- alone regressions. Moreover, while the RMSEs are higher than for the original data (combined data RMSE 2.863), the degradation is marginal: +0.002 (self-reconstruction) and +0.010 (cross-imputation). In line with the previous section, we find that while cross-imputation (i.e., cross-dataset posterior predictives) results in lower goodness-of-fit, it is not at a degree expected to substantially obscure managerial insights. The pattern is more pronounced under differential privacy, as expected, where fused-data RMSE is higher by +0.089 (self-reconstruction) and +0.090 (cross-imputation). Moreover, it is worth noting that the regression metrics from the combined generated data without DP also compare favorably to those of the original individual datasets (RMSE 2.865 vs. survey's 3.093 and CRM's 2.887).

Lastly, we consistently see the same pattern for the anonymous external survey, vis-à-vis original and generated data (Table 1.3). Beyond demonstrating that the fused data generated from PPDF lend greater explanatory and predictive power than either dataset alone (synthetic and original), this exercise also shed light on the inferential consistency between the internal and focal external survey data generating processes. This is an important insight for the partnering firm, as well as for future end-users of our framework. The internal survey represents a valuable, but ultimately costly, one-time, data collection exercise for the telecom carrier. Our evidence here is meant to exemplify PPDF's ability to provide ground-truth consistent results, but future end-users need not have an equivalent 'calibration' to utilize this framework, which is designed directly for fusion of anonymous surveys to other data sources.

Predicting Churn with PPDF.

We now show how combining CRM and survey data can improve our understanding of churn. Churn is a common variable-of-interest that managers seek to predict

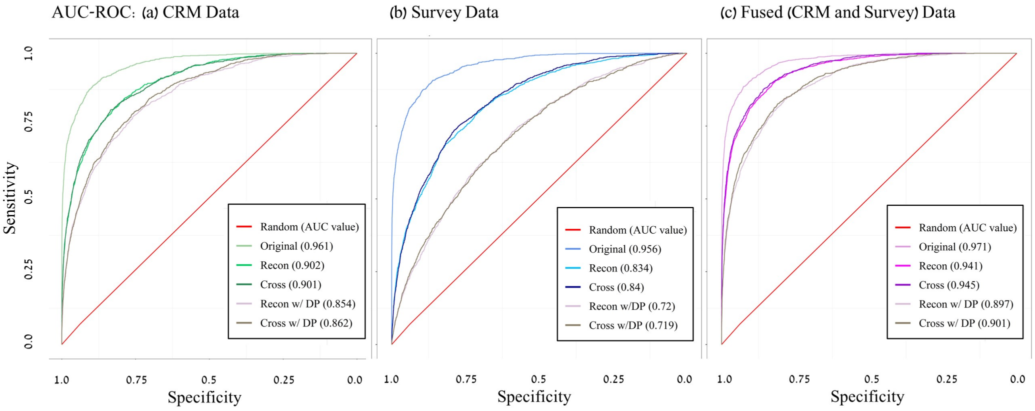


FIGURE 1.9: AUC-ROC curves measuring the accuracy of predicting churn using the CRM data only (a), internal survey data only (b), and fused (combined) data (c). This accuracy measure is based solely on the internal survey because it is the only survey where we can deterministically link churn to the survey response. The best goodness-of-fit comes from the combined data.

and explain. We use a binary logistic regression to study the predictors of churn. This regression uses the data from the survey, which was conducted a year earlier (Q3 2021) than the observed churn (Q3 2022). The challenge here is to incorporate data from the smaller sample of anonymous survey respondents into the larger random sample from the full customer base.

Figure 1.9 presents the AUC-ROC curves of the binary churn model for the internal (calibration) survey respondents (20K). The figure illustrates the trade-off when the churn model is run using the original data vs. the self-reconstructed, cross-imputed, and differentially private data.

The key pattern to observe is the relative position of these specifications' curve. A curve that is pulled more towards the top-left indicates higher accuracy in predicting churn. The purpose of once again utilizing the internal survey lies in our ability to deterministically link the 20K respondents from Q3 2021 to their churn outcome in Q3 2022. As most U.S. telecom carriers utilize a subscription model, churn is defined by explicit termination of contract. Figure 1.9 is organized by data source, i.e., CRM data only (panel a), survey data only (b), and fused (combined) data (c). The original CRM, survey, and combined data lead, unsurprisingly, to the best goodness-of-fit relative to the synthetic datasets arising from the PPDF data fusion, either with or without differential privacy. Even without differential privacy, there is a non-zero loss in accuracy when applying PPDF, which is true for all generative models at the time of writing (Abadi et al., 2016; Bagdasaryan, Poursaeed, and Shmatikov, 2019)

Compared to the AUC of the regression on the original data (CRM 0.961, survey 0.956, and combined 0.971), the loss in accuracy for self-reconstruction (CRM 0.902, survey 0.834, and combined 0.941) and cross-imputation (CRM 0.901, survey 0.840, and combined 0.945) are comparable, and manifest as largely overlapping curves in

Figure 1.9. Moreover, across all models, adding differential privacy negatively impacts respective model predictions, as seen in the DP models' curves being pulled towards the 45-degree "random" line. These results on forward-looking churn reaffirm the findings from the LTR analysis that (1) combing the data sources lead to a gain in predictive accuracy, (2) this phenomenon is true for both the original ground-truth data, as well as the self-reconstructed and cross-imputed fused data, and (3) the pattern holds true under differential privacy, albeit with a trade-off between privacy and accuracy.

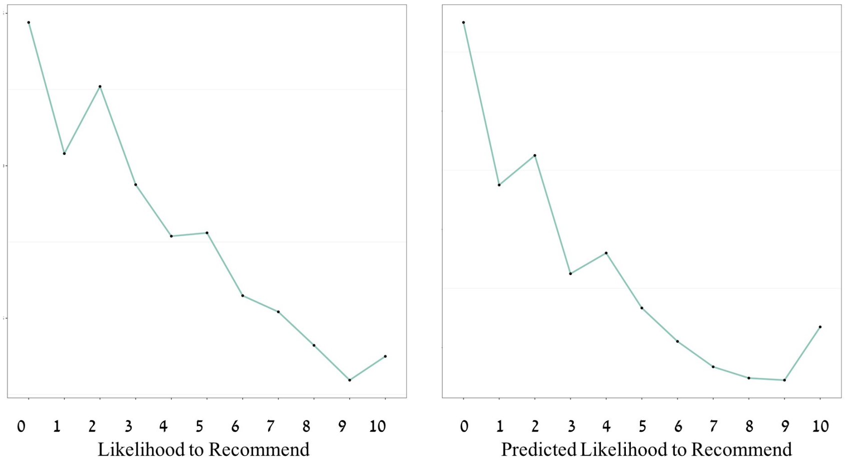


FIGURE 1.10: Left: observed churn (Y-axis redacted for confidentiality) as a function of observed LTR from the internal calibration survey. Right: observed churn of the "full" customer base of 3.6 million randomly selected customers, as a function of predicted LTR, even though these customers did not necessarily respond to the survey.

Customer Base Churn and Satisfaction Predictions.

We now present the results of PPDF in quantifying the relationship between customer churn and satisfaction, the latter measured as LTR. Worth noting, in the absence of a data fusion method, any firm that finds itself in possession of an anonymous customer survey cannot systematically, and in a principled fashion, relate satisfaction and churn outcomes at the individual-level. In the case of the telecom carrier, as we have shown in the previous section, a binary churn model on the (internal calibration) survey responses is improved by utilizing the combined data sources. As such, the ability to generate samples from the posterior predictive distribution of survey responses of the wider customer base is of notable economic significance for these firms. This can potentially enable greater explanatory and predictive power for downstream exercises such as segmentation, CLV models, and churn management. To this end, across this section and the next, we examine and describe the cross-imputation of survey outcomes for the sample of 3.6 million of the telecom carrier's customers. The outcomes can be understood as a counterfactual analysis on how these customers would have responded to the surveys had they chosen to take them. Here, we specifically focus on the relationship between imputed LTR and the forward-looking churn, as defined earlier.

Given that the external survey response rate is <0.01% of the customer base each quarter, the cross-imputation of survey outcomes of the wider customer base is, by definition, a known unknown. Therefore, unlike the exercises above, where we could rely on the calibration survey for customer-level ground-truth validation, here we appeal to aggregate-level correlations. Figure 1.10 illustrates a generally negative relationship between LTR and churn rate (corr=-0.32), aggregated by the eleven LTR bins (0 to 10), for both the internal calibration survey (20K, left panel) and for the cross-imputation to the 3.6 million sample from the customer base (right panel). As a reminder, the internal calibration survey enables direct linkage between LTR and churn via the respondents' user ID. While such direct linkage is evidently nonexistent for the cross-imputed customer base, our ability to recover negative correlation similar to the calibration survey supports the ecological validity of PPDF's imputations.

Importantly, unlike the calibration results, which are limited to a small subset of respondents, managers can undertake proactive churn management based on imputed LTR cutoffs. At a high level, given a retention budget, the carrier can use the imputed LTR values to define a cutoff (i.e., segment) for the most churn-vulnerable customers. However, it's worthwhile to note that the relationship between imputed LTR and churn is not a simple negative linear relationship. Of note, in Figure 1.10 (left and right), for both the ground-truth calibration survey and the imputed customer base, we observe a 'reversal' in churn propensity of customers at the highest LTR (10/10). This is a surprising finding that merits attention. We hypothesize that this is driven by these customers' attractiveness to competitors. In other words, the higher churn in this LTR group may be driven by the "competition vulnerable, high customer lifetime value" segment. This may make those with the highest predicted LTR suitable for interventions that will make them more likely to remain customers (Lemmens and Gupta, 2020).

Selectivity Correction.

In conclusion, although our analyses of specific covariates were focused on LTR and churn, given their outsized importance and prevalence at most direct-to-consumer companies, we wish to emphasize that data fusion was successfully conducted across the entire bench of features from both the customer surveys and the CRM database. Moreover, as noted above, counterfactual customer base cross-imputations were not limited to just LTR. As such, it is easy to imagine that the analysis from Subsection 1.5.3 could be repeated for all of the imputed perception questions.

In lieu of postulating managerial implications of individual perception questions, we end by broadly highlighting the value of data fusion of the anonymous

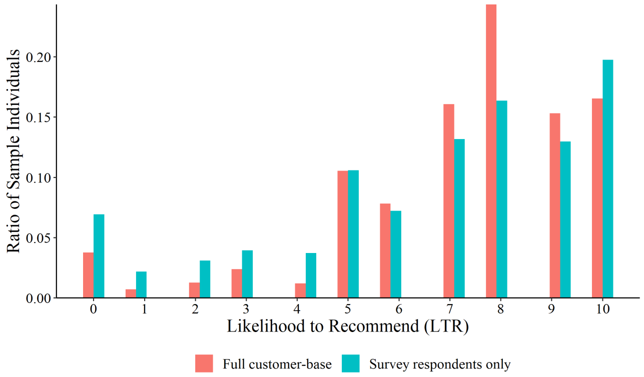


FIGURE 1.11: Predicted LTR of the 3.6 million sample from the full customer base, in red, as imputed by PPDF, along with the observed LTR of the 8K external survey respondents (in blue). The differences highlight the importance of correcting for selection bias before extracting insights from this and similar surveys. Survey respondents are more extreme in their Likelihood to Recommend (along both extremes).

external survey to the wider customer base, in particular, the resulting re-sampling of the survey response distributions. In Figure 1.11, we present the density plots of predicted and observed Likelihood to Recommend. More so, as seen in Figure A.1, in all cases across thirty-four questions, there is notable deviation in the questionnaire response distribution between the external survey as-is (8K respondents) and the customer base cross-imputations (3.6 million). As noted in the discussion on Handling Missingness and Selection Bias, our data fusion framework corrects for selection-on-observables under the assumption that variables that affect the decision to respond to the survey are observed in the fused dataset.

Without data fusion, firms are left without a principled approach to reconciling survey responses, which are known to be biased due to respondent self-selection (Bethlehem, 2010), to a more representative distribution of perception and sentiments of the customer base. Equipped with PPDF, in addition to privacy guarantees, firms need not to limit themselves to just the responses of those who chose to participate, but rather, they can allow the underlying generative framework to provide an unbiased and representative imputation of likely survey outcomes of their overall customer base to enable more robust decision making using the combined information of customer surveys and internal CRM data.

1.6 Summary

In this research, we presented a Privacy Preserving Data Fusion framework and demonstrated its applicability for fusing data sets across various contexts, showcasing the scalability, expressiveness, and generalizability of our approach to future end-users, including a demonstrative case with a large-scale anonymous survey and the CRM database from a leading U.S. telecom carrier.

The challenges of marketing automation and analytics in the era of data privacy are ongoing and multifaceted. This project aims to understand how data fusion, a prevalent marketing analytics technique, can be better retooled to meet today's new privacy standards and practices. Our methodology offers a practical solution to collecting and fusing disparate datasets while protecting consumer anonymity. We show that collecting data while protecting privacy does not mean forgoing the advantages and insights that existing data fusion techniques allow. Using PPDF, companies can safely fuse datasets without them "ever meeting one another", and potentially even split data, thus protecting customers' fundamental right to privacy and reducing the risks associated with data breaches and leaks.

To demonstrate the managerial utility of our methodology, we show how combining CRM and survey data can improve our understanding of churn. We find a generally negative relationship between Likelihood to Recommend (LTR) and churn rate for both the internal calibration survey and for the cross-imputation to the 3.6 million customer random sample of the the telecom carrier's customer base. The internal calibration survey enables direct linkage between LTR and churn via respondent user ID. These, and the results from the fusion between the external survey and sample of full CRM customer base allow us to quantify the relationship between LTR and potential churn. Our methodology enables managers to use LTR as a potential segmentation strategy for proactive churn management while ensuring that customer privacy is preserved.

Our methodology can address different contexts in marketing and beyond, outside the scope of this chapter. We focused our application on inferring missing customer attributes from anonymous surveys. However, further application of PPDF can be used by two or more firms, who wish to learn from the joint distribution of their databases to gather market research insights, such as understanding their market share, or assess complementarity and substitutability in their products, while protecting their intellectual property and the privacy of their customer base.

Furthermore, PPDF can potentially enhance privacy guarantees by splitting a sensitive dataset into two or more datasets. If a data breach were to occur, such separation ensures that data obtained will not be as harmful as the original, joint data. The harms of severe data breaches can be dramatically reduced if names, email addresses, and other identifiers are not stored alongside sensitive choices, attitudinal, and other individual-level data. Firms would be willing to split their data only if, when an insight on the joint data is requested, such insight would be possible in a secured manner, through privacy preserving data fusion. This is possible using PPDF.

The goal of PPDF is to preserve customer anonymity and a company's intellectual property, while enabling the aforementioned use-cases, among others. We exemplify the usage of PPDF to assist a wireless company in exploring the relationships between customer satisfaction and churn. Further research can both expand this framework to investigate causal antecedents and temporal variations of churn from customer perceptions as well as use PPDF in other contexts to answer other substantive questions.

Chapter 2

Understanding Consumer Expenditure Through Gaussian Process Choice Models

joint with Alan Montgomery

Consumers change their choice as expenditures within a category increase. Traditional choice models usually make restrictive structural assumptions to specify the expenditure elasticity. This imposed functional form of utility strongly influences the range of estimable substitution patterns across goods. Consumers with highly nonlinear preferences may have consumption thresholds in which buying patterns dramatically change when price or budget changes. Understanding these thresholds with a flexible utility-based model could lead to improved pricing and promotion decisions. Using Gaussian process priors on utility functions, the functional form on the outside good utility is estimated within the context of constrained utility maximization. In a first application, we estimate a general direct utility choice model for simultaneous purchases within a product category. In a second application, we relax additivity and allow for complementarity in a nonhomothetic choice model, estimating the outside good functional form. Using simulations, our model captures non-linear rates of satiation and precise baseline preferences that traditional non-homothetic parametric models fail to capture by assuming a given functional form of utility. The proposed model automatically detects non-linear patterns of consumption from the data and provide a more precise statistical inference.

2.1 Introduction

Consumers change their choice as expenditures within a category increase. Traditional choice models usually make restrictive structural assumptions to specify the expenditure elasticity. In this chapter, we show that an imposed functional form of utility strongly influences the range of estimable substitution patterns across goods. Consumers with highly nonlinear preferences may have consumption thresholds in which buying patterns dramatically change when price or budget changes. Understanding these thresholds with a flexible utility-based model should lead to improved pricing decisions. For example, consumers with homogeneous preferences who are making a choice within a category between two product tiers: the low quality yogurt tiers with two varieties, a basic vanilla yogurt and a basic strawberry yogurt; and a high quality yogurt tiers with two varieties, a premium vanilla yogurt and a premium strawberry yogurt. Let us suppose that consumers prefer vanilla to strawberry when they decide on their first unit of consumption. Consumers often purchase more than one variety within the sub-category. For example, consumers may purchase simultaneously a vanilla yogurt and a strawberry yogurt, rather than two vanilla yogurts, since the marginal utility of consuming an additional vanilla yogurt is less than the marginal utility of consuming a strawberry yogurt.

Now, we assume that consumers are facing a price reduction for the basic vanilla yogurt, or they budget exogenously increases for a given trip. Will they buy more basic vanilla yogurts? Due to an enlarged budget constraint, consumers may switch to the premium vanilla yogurt and/or to the strawberry yogurt. If we keep utility constant, and if a price reduction occurs in the low quality category product, it is likely that the demand for the low quality product will not increase as much as if they decided to give a price discount on the high category product. This asymmetric switching has been documented in the literature, in notably Allenby and Rossi (1991) and Allenby, Garratt, and Rossi (2010). However, consumers also seek to buy different varieties (strawberry and vanilla) due to satiation when buying a unique variety.

Ideally, a general utility model should be able to accommodate simultaneously normal and superior goods, i.e. allowing non-constant marginal utility functions, and demand for variety i.e. allowing different satiation rates for different goods. Two separate mechanisms are at play: a satiation or substitution effect, and a trading-up or income effect. Consumers become increasingly satiated and consume a different variety within the same sub-category, given a vector of prices and total expenditure. Nevertheless, when expenditure increases and the budget constraint is relaxed (either through an exogenous income increase or through price discounts), consumers start to trade up from the low quality tiers to the high quality one. It has often been assumed that utility for each product bought by consumers is constant. This assumption is overly restrictive as we just saw in our example. Constant utility implicitly entails that consumers should buy the same items as their expenditure increases in the category, whereas trading-up is frequently observed in practice. The literature on non-homothetic choice models has focused on parametric specifications, such as the translog model (e.g. Chiang (1991)), Stone-Geary utility (J. Kim, Allenby, and Rossi, 2002), and rotated indifference curves (Allenby and Rossi, 1991; Allenby, Garratt, and Rossi, 2010). A nonhomothetic choice model is suitable when goods within the category of interest have wide differences in quality or for specifying preferences across categories (Chintagunta and Nair, 2011). Yet, existing nonhomothetic choice models do not capture both variety seeking behavior (driven by satiation) and trading-up (switching from low to high quality-tiers).

We hypothesize that expansion in the category moderates variety seeking. Consumers rapidly reach satiation for a low quality product, and then switch to a higher quality product. Satiation arises because the marginal utility of consuming an extra unit of the same product will be lower than the marginal utility of consuming a unit of a different product. However, satiation may change when consumers start to trade up to the high quality tiers. In our example, the satiation for the premium vanilla yogurt might be much lower than with the basic vanilla yogurt. The consequence of this conjecture is that parametric forms of utility functions used in the literature display undesirable properties that limit the range of estimable consumption patterns.

The previous limitations in traditional economic models of choice leads us to conjecture that the assumption that the random utility follows a parametric specification is simplistic. Parametric assumptions on the utility function can induce potentially large errors when the parameterization is misspecified (Gu, Bhattacharjya, and Subramanian, 2018). Moreover, marketers may have reasons to anticipate highly non-linear preferences in some of the goods. Failing at modeling these non-linearities would prevent marketers from exploiting thresholds effects in consumption, and would translate into a missed opportunity to increase profit. A general utility model should be able to automatically detect non-linear patterns of consumption from the data. Yet, no model in the marketing literature has used a general utility specification to estimate demand systems for simultaneous purchases, related to the approach of Wales and Woodland (1983) and Hanemann (1984). A corollary of that issue is the need for a principled way of modeling the uncertainty around the shape of utility functions inferred in a nonparametric choice model.

Gaussian Processes (GPs) are a popular tool for nonparametric function estimation; they have numerous desirable properties, as they automatically produce estimates on prediction uncertainty, work as interpolators for a consistent decision maker, and retain their Gaussian property when transformed by linear operators. Moreover, GPs are also able to encode prior knowledge in a principled way. Marketing research has recently featured Gaussian processes to capture the individual-level dynamics of heterogeneity and its implications for targeting, pricing and market structure analysis (Dew, Ansari, and Y. Li, 2020). Gaussian processes can also model latent functions that determine purchase propensity, and can be used to analyze purchasing dynamics in which known and unknown calendar time determinants of purchasing with individual-level predictions are joined together (Dew and Ansari, 2018).

We build a general, flexible random utility framework to estimate demand for simultaneous purchases within and across product categories. Using Gaussian process priors on utility functions, we relax the functional form on the outside good, within the context of constrained utility maximization. In our first application, we derive a demand system where the logarithm of the first-order derivatives of the utility function follow a Gaussian Process. We show in numerical simulations that misspecified functional form of outside good utility leads to inconsistent estimation of preference parameters on the inside goods. We demonstrate that our proposed framework is robust to misspecification and enable the modeler to recover the true preference parameters in the category of interest. In our second application, we relax the additivity assumption and propose a nonhomothetic choice model that can accommodate complementary goods, using household production theory. We show in numerical simulations, that the misspecified functional form of outside good utility also leads to inconsistent estimation of preference parameters on each good. Similarly, our proposed Gaussian process framework is robust to misspecification in this context and accurately recovers the true preference parameters.

We estimate a general direct utility choice model the no-U-turn sampler (NUTS) variant of Hamiltonian Monte Carlo. Our model captures non-linear rates of satiation for inside and outside goods alike, that traditional non-homothetic parametric models fail to capture by assuming a given functional form of utility. The proposed model can automatically detect non-linear patterns of consumption from the data and provide a more precise statistical inference, since it flexibly models consumer behavior and achieve a better fit than a parametric model that assumes specific functional forms on the sub-utility functions.

The proposed framework detects rich patterns in the data, especially non - linearities in preferences that could not be previously captured with less flexible models with specific utility functions such as in J. Kim, Allenby, and Rossi (2002) and Allenby and Rossi (1991). Pricing products using a utility model with a misspecified functional form of utility has negative consequences on the firm profit. If the true rate of satiation is much higher than the estimated rate using parametric form of utility, then price reductions on these items arising from the estimated rate of satiation are sub-optimal since consumers will not buy from the items that display high satiation rates. Due to its structural nature, the model is able to provide us with counterfactual predictions on pricing, promotion and expenditure elasticities.

2.2 Methodological Background

2.2.1 Literature on Non-Homothetic Choice and Demand Modeling

The literature on non-homothetic choice models has focused on parametric specifications, such as the translog model (e.g. Chiang (1991)), Stone-Geary utility (J. Kim, Allenby, and Rossi, 2002), and rotated indifference curves (Allenby and Rossi, 1991; Allenby, Garratt, and Rossi, 2010). A nonhomothetic choice model is suitable when goods within the category of interest have wide differences in quality or for specifying preferences across categories (Chintagunta and Nair, 2011). Yet, existing nonhomothetic choice models do not capture both variety seeking behavior (driven by satiation) and trading-up (switching from low to high quality-tiers).

Allenby and Rossi (1991) and Allenby, Garratt, and Rossi (2010) use an implicitly defined utility function marginal utility. However, their model assumes linear indifference curves, and rules out multiple discreteness, which limits its empirical use. Allenby and Rossi (1991) were able to capture non-homothetic preferences with their multinomial probit. But consumers only select one brand at a time and for only one unit, so no variety is allowed. Allenby, Garratt, and Rossi (2010) mostly focus on capture how advertising affect the rate at which consumers are willing to trade up to higher quality brands, but its structural multinomial logit specification does not incorporate multiple purchase incidence.

J. Kim, Allenby, and Rossi (2002) allow for multiple brands and units to be purchased simultaneously, and focus on consumer demand for variety, with non - homothetic preferences. They propose a horizontally differentiated demand model based on a translated additive utility structure, while allowing for the possibility of a mixture of corner and interior solutions where more than one but not all varieties are selected. They use the following utility parametrization:

$$U(q_1, ..., q_J) = \sum_{j}^{J} \psi_j (q_j + \gamma_j)^{\alpha_j} \tag{2.1}$$

where ψ_j is the baseline utility parameter, α_j is the parameter that ensures diminishing marginal returns and γ_j is a location translation parameter that translates the utility function to accommodate both interior and corner solutions. When $\gamma_j = 0$, only interior solutions are allowed for good j. When a good has a large baseline utility and a value of α_j close to one, purchases of large quantities of only one variety (high baseline preference and low satiation, and small values of α imply a high-satiation rate. However, J. Kim, Allenby, and Rossi (2002) pointed out that the difficulty to separately identify α and γ since both parameters govern the slope of the indifference curves at the point of intersection with the axes, and they have to fix γ to 1 for all goods. Fixing this critical parameter for all goods shows the limitations of this model, as the patterns of substitution across goods are then substantially restricted. When $\alpha_j < 1$, consumers' marginal utility diminishes with increased consumption; consumer are then satiated and pushed toward multiple discreteness. However, the superior or inferior nature of the good is confounded with variety seeking. Both α and ψ influence the rate of satiation as both parameter in the second derivative of the utility function, which clearly indicates a deficiency in the parametric form. Finally, this utility specification rules out the possibility of utility to be bounded, including rapid satiation rates inducing plateaus of consumption, or sudden changes of regimes after a certain amount of good is consumed. Furthermore, the literature has proposed alternative explanations, such as time varying preferences (Hasegawa, Terui, and Allenby, 2012), in which product attributes and dynamic effects are incorporated in the baseline utility and satiation parameters. However, time-varying preferences is not needed to capture nonlinearities in preferences, and as such may lack parsimony. Moreover, Hasegawa, Terui, and Allenby (2012) assume a logarithmic parametric form for the outside good, which is still restrictive since the price effects only depend on the preferences of the inside goods and not the demand of the outside good.

2.2.2 Gaussian processes as Priors on Latent Functions

A Gaussian process (GP), denoted as a stochastic function $f(\cdot)$, is established over a domain of interest, which, for our purposes, is represented by quantity $q \in \mathbb{R}^+$. The defining characteristics of a GP are its mean function $m(q)$ and covariance function $k(q, q')$, where, given a specific set of input times $q = q_1, q_2, \ldots, q_T$, the function values are distributed as $f(q) \sim \mathcal{N}(m(q), K(q))$. Here, $m(q)$ stands for the mean function evaluated across all inputs, yielding a $T \times 1$ vector, and $K(q)$ is a $T \times T$ covariance matrix, constructed by pairwise evaluation of the covariance function $k(q, q')$ across the inputs. Briefly, the mean function establishes the prior mean of the process's value for each quantity input q, while the covariance function delineates the extent of correlation between the process values at different quantity pairs q and q'. Given that a GP configures a probability distribution over potential outputs for any given set of inputs, it inherently provides a flexible, nonparametric prior over latent function spaces, an aspect critically useful in Bayesian data analysis (Rasmussen, Williams, et al., 2006), and is typically represented as $f(q) \sim \text{GP}(m(q), k(q, q'))$.

In the realm of GP applications, mean functions are often considered secondary and commonly presumed constant. This assumption allows the covariance function, or kernel, to describe the essential characteristics of the functions delineated by the GP priors. These kernels can encapsulate various general traits of the modeled functions, like smoothness, differentiability, and amplitude. A kernel, denoted by $k : \mathbb{R}^2 \to \mathbb{R}$, ensures the generated covariance matrix $K(q)$ is positive semidefinite across any inputs q. The GP literature introduces numerous kernels, with the squared exponential (SE) kernel being the simplest and most favored. The SE kernel, chosen for our discussion, is parameterized by an signal noise parameter σ and a lengthscale parameter β, described by the formula:

$$K_{SE}(q, q' | \sigma^2, \beta) = \sigma^2 \exp\left(-\frac{(q - q')^2}{2\beta^2}\right) \tag{2.2}$$

The signal variance parameter σ^2 determines the potential deviation of function values from the mean, while the lengthscale parameter β accounts for the smoothness of these deviations, also referred to as the smoothness parameter. The straightforward yet potent nature of the SE kernel facilitates its broad application in prior marketing research (Dew and Ansari, 2018; Dew, Ansari, and Y. Li, 2020; Dew and Fan, 2021; Dew, Ascarza, et al., 2023), emphasizing its capacity to impose priors on latent functions.

2.2.3 Flexibly Modeling Quality Tiers Effects

Nonhomothetic preferences represent a departure from traditional utility models, recognizing that the proportion of income spent on different goods can vary across income levels. This perspective is crucial for understanding consumer behavior related to trading up, where consumers opt for higher-quality—and often more expensive—products as their income increases. The literature suggests that such preferences can significantly influence market dynamics, including product positioning and the competitive landscape.

Traditional approaches to modeling quality tiers often rely on restrictive assumptions about the functional form of utility derived from the consumption of an outside good. This sensitivity to functional form assumptions can lead to models that either underestimate or overestimate the attractiveness of trading up or down between quality tiers. The choice of utility function significantly impacts the inferred elasticity of substitution between goods (C. Kim et al., 2023), which in turn affects predictions about consumer response to price changes, product improvements, or the introduction of new products. The critical takeaway is that the modeling of quality tiers requires careful consideration of these assumptions to ensure accurate representation of consumer behavior.

In response to these challenges, our proposed approach introduces flexibility in modeling quality tiers by allowing for switching between complementarity and substitution effects driven by expenditure. Our approach acknowledges that consumers' preferences between higher and lower-quality tiers can vary based on their expenditure levels, potentially shifting from substitution to complementarity (or vice versa) as their budget constraints change. Such a framework is more aligned with the empirical observations of consumer behavior, capturing the nuances of trading up or down in response to changes in income, prices, or product attributes.

2.2.4 Outside Good in Choice Modeling

C. Kim et al. (2023) investigates the impact of outside good utility functions on substitution patterns within multiple discrete/continuous demand models. The authors present novel results on the functional form of quantity price effects in these models, highlighting the limitations of standard outside good utility functions. A new, more flexible outside good utility function, from the class of hyperbolic functions, is proposed to accommodate broader substitution patterns and address issues of satiation.

In their paper, the utility has the following form:

$$u_z(z) = \frac{\psi_z}{b} \frac{1 + \exp(bz + c)}{1 - \exp(bz + c)} \tag{2.3}$$

with slope parameter $b > 0$ and an intercept parameter $c > 0$. An empirical analysis using household scanner data from the potato chip market supports the model's ability to capture non-standard satiation rates for the outside good, impacting price elasticity estimates and the effectiveness of loyalty coupon targeting programs. Their research contributes to understanding the role of outside good utility in direct utility models, proposing a more adaptable approach to modeling consumer choice behavior. However, this utility function is still parametric and impose a functional form, which impose strong restrictions on substitution patterns across inside goods.

2.3 Application 1: Within-Category Nonhomothetic Demand Model

2.3.1 Random Utility Model

We are interested in performing inference on the functional form of the marginal utility for a given consumer. At a given purchase occasion, we assume that consumer's choose a vector of quantities $\boldsymbol{q} = (q_1, ..., q_J)$ that maximize their direct utility function, whose support is the vector of quantities consumed. We suppress the time t and individual i subscript in this section for the moment and will introduce them later. Our utility model is strongly separable across a vector of inside products at our focal store ($j = 1, 2, ..., J$). We assume that the sub-utility of the good j, $U_j(q_j)$, and the sub-utility of the outside good $\nu_z(z)$ are continuously differentiable, increasing functions. Following the standard random utility approach, we introduce a multiplicative error term into our utility model:

$$U(q_1, ..., q_J) := \sum_j u_j(q_j) := \sum_j \left(\nu_j(q_j) \exp\left(\varepsilon_j\right)\right) \tag{2.4}$$

where, for all $j = 1, \ldots, J$, the error terms ε_j are i.i.d. and follow a Type-1 extreme value distribution with scale σ fixed, for example to unity. The random error ε_j is known by consumers but unobserved by the researcher. Moreover, ε_j and ν_j the sub-utility function for the good j are independent. In addition, ε_j and $\frac{\partial \nu_j}{\partial q_j}$, the marginal sub-utility function for the good j are also independent. We also assume that the outside good is always consumed, such that $z > 0$.

This random element represents information that is known to consumers, but not observed by the researcher. This yields a form in which the marginal utilities consist of a deterministic $\frac{\partial v_j}{\partial q_j}$ and a random ε_j component:

$$\frac{\partial U_j(q_j)}{\partial q_j} = \frac{\partial v_j(q_j)}{\partial q_j} \exp\left(\varepsilon_j\right) \tag{2.5}$$

which yields:

$$\log\left(\frac{\partial U_j(q_j)}{\partial q_j}\right) = \log\left(\frac{\partial v_j(q_j)}{\partial q_j}\right) + \varepsilon_j \tag{2.6}$$

The logarithmic form ensures that the marginal utility remains positive. The budget constraint is linear. Consumer maximize utility in a static fashion, such that there is no forward-looking behavior or savings allowed.

$$\sum_j p_j q_j = x \tag{2.7}$$

where p_j is the price of good j, and x represents the total expenditure. The corresponding utility maximization problem may be written in Lagrangian form:

$$\max_{\boldsymbol{q},\lambda} \mathcal{L} = U(\boldsymbol{q}) - \lambda\left(\sum_j p_j q_j - x\right) \tag{2.8}$$

where $\lambda > 0$, the Lagrange multiplier of the utility maximization problem. After dividing by prices and taking logs, the Kuhn-Tucker first-order conditions are:

$$\log\left(\frac{\partial v_j(q_j)}{\partial q_j}\right) + \varepsilon_j - \log(p_j) = \log(\lambda) \qquad \text{if } j \text{ s.t. } q_j > 0 \tag{2.9}$$

$$\log\left(\frac{\partial v_j(q_j)}{\partial q_j}\right) + \varepsilon_j - \log(p_j) < \log(\lambda) \qquad \text{if } j \text{ s.t. } q_j = 0 \tag{2.10}$$

Without loss of generality, we can assume that the first good is always purchased, and we take the difference of the first-order conditions. This differencing ensures that the budget constraint is satisfied (J. Kim, Allenby, and Rossi, 2002). Then the Kuhn-Tucker conditions can be rewritten as follows:

$$\varepsilon_j = V_1 - V_j + \varepsilon_1 \qquad \text{if } j \text{ s.t. } q_j > 0 \tag{2.11}$$

$$\varepsilon_j < V_1 - V_j + \varepsilon_1 \qquad \text{if } j \text{ s.t. } q_j = 0 \tag{2.12}$$

where $V_j = \log\left(\frac{\partial v_j(q_j)}{\partial q_j}\right) - \log(p_j)$ for all $j = 1, ..., J$.

Note that the Kuhn-Tucker first order conditions are necessary and sufficient when the utility function is monotonic and strictly quasi-concave (Dubé, 2019). Quasi-convexity of the constraint function is also needed but it is automatically true since the constraint function is linear.

Let the joint probability density function of the ε_k terms be $f(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_J)$. The likelihood can be constructed following C. R. Bhat (2005) and C. R. Bhat (2008) by showing that the probability that a consumer allocates on the occasion t (but omitting the subscript) all expenditure to the first M of the J goods at the optimum (denoted by q_j^* for all $j = 1, \ldots, J$) is:

$$L(q_1^* > 0, q_2^* > 0, \ldots, q_M^* > 0, q_{M+1}^* = 0, \ldots, q_J^* = 0) \tag{2.13}$$

$$= |\boldsymbol{J}| \int_{\varepsilon_1=-\infty}^{\varepsilon_1=+\infty} \int_{\varepsilon_{M+1}=-\infty}^{V_1-V_{M+1}+\varepsilon_1} \int_{\varepsilon_{M+2}=-\infty}^{V_1-V_{M+2}+\varepsilon_1} \cdots \int_{\varepsilon_J=-\infty}^{V_1-V_J+\varepsilon_1}$$

$$f(\varepsilon_1, V_1 - V_2 + \varepsilon_1, V_1 - V_3 + \varepsilon_1, \ldots V_1 - V_M + \varepsilon_1, \varepsilon_{M+1}, \varepsilon_{M+2}, \ldots, \varepsilon_{J-1}, \varepsilon_J)$$

$$d\varepsilon_J d\varepsilon_{J-1}, \ldots, d\varepsilon_1 \tag{2.14}$$

where $\boldsymbol{J}$ is the Jacobian matrix whose elements are given by:

$$\boldsymbol{J}_{kl} = \frac{\partial\left(V_1 - V_{k+1} + \varepsilon_1\right)}{\partial q_{l+1}} \tag{2.15}$$

$$= \frac{\partial\left(V_1 - V_{k+1}\right)}{\partial q_{l+1}} \qquad \text{for } k, l = 1, 2, \ldots, M-1 \tag{2.16}$$

Using Type-1 extreme value distribution on the error term, we can write the closed form likelihood:

$$L(q_1^*, q_2^*, \ldots, q_M^*, 0, \ldots, 0) = \frac{1}{\sigma^{M-1}} \left(\prod_{j=1}^{M} \left| -\frac{\partial}{\partial q_j} \log\left(\frac{\partial v_j}{\partial q_j}\right) \right| \right) \left(\frac{1}{p_1} \left| \sum_{j=1}^{M} \frac{p_j}{-\frac{\partial}{\partial q_j} \log\left(\frac{\partial v_j}{\partial q_j}\right)} \right| \right) \tag{2.17}$$

$$\left(\frac{\prod_{j=1}^{M} \exp\left(\frac{V_j}{\sigma}\right)}{\left(\sum_{j=1}^{J} \exp\left(\frac{V_j}{\sigma}\right)\right)^M} \right) (M-1)!$$

where $V_j = \log\left(\frac{\partial v_j}{\partial q_j}\right) - \log(p_j)$ for all $j = 1, ..., J$. The full derivation of the likelihood is in Appendix B.1.

2.3.2 Vectorized notation and functional notation

For the remainder of the chapter, we also adopt the following vectorized notation for all occasions $t = 1, ..., T$ at the individual level. Let $\boldsymbol{q}_j = \left[q_{j1}, \ldots, q_{jT}\right]$ be the vector of optimal quantity j for all occasions, $\boldsymbol{p}_j = \left[p_{j1}, \ldots, p_{jT}\right]$, the vector of prices j for all occasions, $\boldsymbol{U}_j(\boldsymbol{q}_j) = \left[U_j(q_{j1}), \ldots, U_j(q_{jT})\right]$, the vector of stochastic utility j evaluated at the optimal quantity j for all occasions, $\boldsymbol{\nu}_j(\boldsymbol{q}_j) = \left[\nu_j(q_{j1}), \ldots, \nu_j(q_{jT})\right]$, the vector of deterministic utility j evaluated at the optimal quantity j for all occasions, $\frac{\partial \boldsymbol{\nu}_j(\boldsymbol{q}_j)}{\partial \boldsymbol{q}_j} = \left[\frac{\partial \nu_j(q_{j1})}{\partial q_j} \ldots, \frac{\partial \nu_j(q_{jT})}{\partial q_j}\right]$, the vector of deterministic marginal utility j

evaluated at the optimal quantity j for all occasions, $\frac{\partial^2 \boldsymbol{\nu}_j(\boldsymbol{q_j})}{\partial \boldsymbol{q}_j^2} = \left[\frac{\partial^2 \nu_j(q_{j1})}{\partial q_j^2} \ldots, \frac{\partial^2 \nu_j(q_{jT})}{\partial q_j^2}\right]$, and $\boldsymbol{\varepsilon}_j = \left[\varepsilon_{j1}, \ldots, \varepsilon_{jT}\right]$ is the vector of error terms for utility j for all occasions.

We also adopt a functional notation, and use $\boldsymbol{\nu}_j$, $\frac{\partial \boldsymbol{\nu}_j}{\partial \boldsymbol{q}_j}$ and $\frac{\partial^2 \boldsymbol{\nu}_j}{\partial \boldsymbol{q}_j^2}$ to denote respectively the deterministic utility function, the deterministic marginal utility function, and the second derivative of the deterministic utility function. Likewise, $\log\left(\frac{\partial \boldsymbol{\nu}_j}{\partial \boldsymbol{q}_j}\right)$ and $\frac{\partial}{\partial \boldsymbol{q}_j} \log\left(\frac{\partial \boldsymbol{\nu}_j}{\partial \boldsymbol{q}_j}\right)$ denote respectively the logarithm of the marginal utility function and the derivative of the logarithm of the marginal utility function.

Let L_t be the likelihood of a consumer's purchase for m_t alternatives at occasion t. To evaluate the log-likelihood function

$$LL\left(\boldsymbol{q}_1, ..., \boldsymbol{q}_J \middle| \frac{\partial \boldsymbol{\nu}_j}{\partial \boldsymbol{q}_j}_{j=1,...,J}, \frac{\partial^2 \boldsymbol{\nu}_j}{\partial \boldsymbol{q}_j^2}_{j=1,...,J}, \boldsymbol{p}_1 \ldots, \boldsymbol{p}_J, \sigma\right) = \sum_{t=1}^{T} \log(L_t) \tag{2.18}$$

we need to construct a prior distribution for not only for the marginal utility function for good j, $\frac{\partial \boldsymbol{\nu}_j}{\partial \boldsymbol{q}_j}$ but also the second derivative of the utility function for good j, $\frac{\partial^2 \boldsymbol{\nu}_j}{\partial \boldsymbol{q}_j^2}$ for all $j = 1, ..., J$. But due to our reparameterization in terms of the logarithm of the marginal utility function $\log\left(\frac{\partial \boldsymbol{\nu}_j}{\partial \boldsymbol{q}_j}\right)$ and the derivative of the logarithm of the marginal utility function $\frac{\partial}{\partial \boldsymbol{q}_j} \log\left(\frac{\partial \boldsymbol{\nu}_j}{\partial \boldsymbol{q}_j}\right)$, it is more natural to place a prior on these last two measures. A consequence of this reparameterization is that utility for each good will always be increasing with quantities purchased.

2.3.3 Gaussian Process Priors

We model the preferences of a consumer who has T purchase occasions. Preferences are fixed and constant over time, but are latent, and the researcher is only able to observe the quantities purchased for all inside and outside goods, and their corresponding prices. We propose to infer the form of utility using Gaussian processes. Specifically, we place a Gaussian process prior on the logarithm of the deterministic marginal utility for each good $j = 1, ..., J$:

$$\log\left(\frac{\partial \boldsymbol{\nu}_j}{\partial \boldsymbol{q}_j}\right) \sim \mathcal{GP}\left(\boldsymbol{\mu}_j, \boldsymbol{K}_j\right) \tag{2.19}$$

where we have the mean and covariance function defined as follows:

$$\boldsymbol{\mu}_j(\boldsymbol{q}_j) = \mathbb{E}\left[\log\left(\frac{\partial \boldsymbol{\nu}_j(\boldsymbol{q}_j)}{\partial \boldsymbol{q}_j}\right)\right] \tag{2.20}$$

$$\boldsymbol{K}_j(\boldsymbol{q}_j, \boldsymbol{q}_j') = \mathbb{E}\left[\left(\log\left(\frac{\partial \boldsymbol{\nu}_j(\boldsymbol{q}_j)}{\partial \boldsymbol{q}_j}\right) - \mu(\boldsymbol{q}_j)\right)\left(\log\left(\frac{\partial \boldsymbol{\nu}_j(\boldsymbol{q}_j')}{\partial \boldsymbol{q}_j'}\right) - \mu(\boldsymbol{q}_j')\right)\right] \tag{2.21}$$

Thereafter we remove the subscript j on the kernel $\boldsymbol{K}$ for simplicity. When we realize our Gaussian processes to the vectors of optimal quantities $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_J$ chosen by consumers, by marginalization property, we obtain that the T values of the logarithm of the marginal utility function for good j has a multivariate normal prior distribution, with covariance defined by the kernel $\boldsymbol{K}$:

$$\log\left(\frac{\partial \boldsymbol{\nu}_j(\boldsymbol{q}_j)}{\partial \boldsymbol{q}_j}\right) := \left[\log\left(\frac{\partial \nu_j(q_{j1})}{\partial q_j}\right), \ldots, \log\left(\frac{\partial \nu_j(q_{jT})}{\partial q_j}\right)\right]^T \sim \mathcal{N}\left(\boldsymbol{\mu}_j(\boldsymbol{q}_j), \boldsymbol{K}(\boldsymbol{q}_j, \boldsymbol{q}_j)\right) \tag{2.22}$$

which can be rewritten with as a vector of T realized log-marginal utilities at the optimal quantities:

$$\begin{pmatrix} \log\left(\frac{\partial \nu_j(q_{j1})}{\partial q_j}\right) \\ \log\left(\frac{\partial \nu_j(q_{j2})}{\partial q_j}\right) \\ \vdots \\ \log\left(\frac{\partial \nu_j(q_{jT})}{\partial q_j}\right) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_j(q_{j1}) \\ \mu_j(q_{j2}) \\ \vdots \\ \mu_j(q_{jT}) \end{pmatrix}, \begin{pmatrix} K(q_{j1}, q_{j1}) & K(q_{j1}, q_{j2}) & \ldots & K(q_{j1}, q_{jT}) \\ K(q_{j2}, q_{j1}) & K(q_{j2}, q_{j2}) & \ldots & K(q_{j2}, q_{jT}) \\ \vdots & \vdots & \ddots & \vdots \\ K(q_{jT}, q_{j1}) & K(q_{jT}, q_{j2}) & \ldots & K(q_{jT}, q_{jT}) \end{pmatrix}\right) \tag{2.23}$$

where $\boldsymbol{K}(\boldsymbol{q}_j, \boldsymbol{q}_j) = (K(q_{jt}, q_{jt'}))$ is a $T \times T$ matrix with each element $K(q_{jt}, q_{jt'})$ and μ_j is a T dimensional mean vector parameter for the Gaussian process. Notice our GP looks like a regression model in function-space view (see Rasmussen, Williams, et al. (2006), §2.3). We use the squared exponential (SE) covariance function for the kernel:

$$K(q_{jt}, q_{jt'}) = \sigma_f^2 \exp\left(-\frac{1}{2\beta^2}(q_{jt} - q_{jt'})^2\right) \tag{2.24}$$

where β denotes the characteristic length-scale and σ_f^2 is the signal variance, hyperparameters of the GP model. The squared exponential covariance function implies that the covariance is almost unity (when $\sigma_f^2 = 1$) between variables whose corresponding inputs are very close (Rasmussen, Williams, et al., 2006)[1]

We express the joint prior distribution of the log marginal utility $\log\left(\frac{\partial \circ_j}{\partial \mathbf{q}_j}\right)$ and its derivative, i.e. the ratio of the derivative of the marginal utility to the marginal

[1] It has been shown that the squared exponential covariance function corresponds to a Bayesian linear regression model with an infinite number of basis function (*ibid*), which emphasizes the flexibility of such a function. Another advantage of this covariance function is that it is infinitely differentiable, which is useful in our context. Finally, we also choose this function for the interpretability of its hyperparameters. The characteristics length-scale parameter β controls the amount of information that the Gaussian process will borrow around the test input point that needs to be evaluated. A large β means that more information is borrowed, which will smooth the utility function; conversely, a smaller β will make the utility function more prone to capture non-linearities in preferences and rationalize them. The characteristic length-scale β and the signal variance σ_f^2 are weakly identified and their proportion is more important to the predictive performance than their individual value for the Matérn class of covariance function (Diggle, Tawn, and Moyeed, 1998; H. Zhang, 2004), to which belongs the squared exponential covariance function.

utility $\frac{\partial}{\partial \boldsymbol{q}_j} \log\left(\frac{\partial \boldsymbol{\nu}_j}{\partial \boldsymbol{q}_j}\right)$ as a Gaussian process. Since the differentiation operator is a linear operator, the derivative of a Gaussian process is another Gaussian process (Rasmussen, Williams, et al., 2006). Therefore, $\frac{\partial}{\partial \boldsymbol{q}_j} \log\left(\frac{\partial \boldsymbol{\nu}_j}{\partial \boldsymbol{q}_j}\right)$, i.e. the derivative function of the log-marginal utility of good j, also follows a Gaussian process prior, and the joint distribution $\left(\log\left(\frac{\partial \boldsymbol{\nu}_j}{\partial \boldsymbol{q}_j}\right), \frac{\partial}{\partial \boldsymbol{q}_j} \log\left(\frac{\partial \boldsymbol{\nu}_j}{\partial \boldsymbol{q}_j}\right)\right)$ also follows a Gaussian process prior. Following X. Wang and Berger (2016), we can derive the kernels associated with the joint Gaussian process for the log-marginal utility of good j and its derivative function, for $j \in 1, ..., J$. This joint measure also has a Gaussian process prior:

$$\begin{bmatrix} \log\left(\frac{\partial \boldsymbol{\nu}_j}{\partial \boldsymbol{q}_j}\right) \\ \frac{\partial}{\partial \boldsymbol{q}_j} \log\left(\frac{\partial \boldsymbol{\nu}_j}{\partial \boldsymbol{q}_j}\right) \end{bmatrix} \sim \mathcal{GP}\left(\begin{bmatrix} \boldsymbol{\mu}_j \\ \frac{\partial}{\partial \boldsymbol{q}_j}\boldsymbol{\mu_j} \end{bmatrix}, \begin{bmatrix} \boldsymbol{K} & \boldsymbol{K^{01}} \\ \boldsymbol{K^{10}} & \boldsymbol{K^{11}} \end{bmatrix}\right) \tag{2.25}$$

By marginalization property, the joint realization of the log-marginal utility for good j at the optimal, observed vector of quantities for good j over T^* purchase occasions, and its derivative function, have a multivariate normal prior distribution:

$$\begin{bmatrix} \log\left(\frac{\partial \boldsymbol{\nu_j}(\boldsymbol{q_j})}{\partial \boldsymbol{q}_j}\right) \\ \frac{\partial}{\partial \boldsymbol{q}_j} \log\left(\frac{\partial \boldsymbol{\nu}_j(\boldsymbol{q}_j)}{\partial \boldsymbol{q}_j}\right) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}(\boldsymbol{q}_j) \\ \frac{\partial \boldsymbol{\mu}(\boldsymbol{q}_j)}{\partial \boldsymbol{q}_j} \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}(\boldsymbol{q}_j, \boldsymbol{q}_j) & \boldsymbol{K^{01}}(\boldsymbol{q}_j, \boldsymbol{q}_j) \\ \boldsymbol{K^{10}}(\boldsymbol{q}_j, \boldsymbol{q}_j) & \boldsymbol{K^{11}}(\boldsymbol{q}_j, \boldsymbol{q}_j) \end{bmatrix}\right) \tag{2.26}$$

with $\boldsymbol{K}(\boldsymbol{q}_j, \boldsymbol{q}_j) = (K(q_{jt}, q_{jt'}))$ is a $T \times T$ matrix with each element $K(q_{jt}, q_{jt'})$;

$$\boldsymbol{K^{10}}(\boldsymbol{q}, \boldsymbol{q}) = (K^{10}(q_{jt}, q_{jt'})) \tag{2.27}$$

is a $T \times T$ matrix with each element $K^{10}(q_{jt}, q_{jt'})$;

$$\boldsymbol{K^{10}}(\boldsymbol{q}_j, \boldsymbol{q}_j) = \boldsymbol{K^{01}}(\boldsymbol{q}_j, \boldsymbol{q}_j)^T \tag{2.28}$$

and

$$\boldsymbol{K^{11}}(\boldsymbol{q}_j, \boldsymbol{q}_j) = (K^{11}(q_{jt}, q_{jt'})) \tag{2.29}$$

is a $T \times T$ matrix with each element $K^{11}(q_{jt}, q_{jt'})$ such that

$$K^{10}(q_{jt}, q_{jt'}) = \frac{\partial}{\partial q_{jt}} K(q_{jt}, q_{jt'}) = \sigma_f^2 \exp\left(-\frac{1}{2\beta^2}(q_{jt} - q_{jt'})^2\right)\left(-\frac{1}{\beta^2}(q_{jt} - q_{jt'})\right) \tag{2.30}$$

$$K^{01}(q_{jt'}, q_{jt}) = \frac{\partial}{\partial q_{jt'}} K(q_{jt}, q_{jt'}) = \sigma_f^2 \exp\left(-\frac{1}{2\beta^2}(q_{jt'} - q_{jt})^2\right)\left(\frac{1}{\beta^2}(q_{jt'} - q_{jt})\right) \tag{2.31}$$

$$K^{11}(q_{jt}, q_{jt'}) = \frac{\partial^2}{\partial q_{jt} \partial q_{jt'}} K(q_{jt}, q_{jt'}) = \sigma_f^2 \exp\left(-\frac{1}{2\beta^2}(q_{jt} - q_{jt'})^2\right)\frac{1}{\beta^2} \left(1 - \frac{1}{\beta^2}(q_{jt} - q_{jt'})^2\right) \tag{2.32}$$

2.3.4 Identification

Inferences about utility are made through the choices that consumers make. However, utility is not directly observed. As a consequence, our metric of utility cannot be identified uniquely from the data. To illustrate this problem, consider a transformation of utility, $F(U(\boldsymbol{q}))$, that is monotonically increasing, $\frac{\partial F}{\partial U} > 0$. Notice that our Lagrangian can be rewritten using this transform without altering the solution:

$$\mathcal{L} = F(U(\boldsymbol{q})) + \lambda(x - \boldsymbol{p}'\boldsymbol{q}) \tag{2.33}$$

where q denotes here the vector of quantities (outside good included) and p denotes the vector of prices. Since we have assumed that F is monotonically increasing, we can rewrite the first order condition as:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{q}} = \frac{\partial F}{\partial U}\frac{\partial U}{\partial \boldsymbol{q}} - \lambda \boldsymbol{p} = 0 \implies \frac{\partial \mathcal{L}}{\partial \boldsymbol{q}} = \frac{\partial U}{\partial \boldsymbol{q}} - \frac{\lambda}{\frac{\partial F}{\partial U}}\boldsymbol{p} = 0 \tag{2.34}$$

Both $\frac{\partial F}{\partial U}$ and λ are scalars. In the transformed space, we can think about our Lagrange multiplier being scaled by our utility transformation. This points out the dependence of utility scale on the Lagrangian. Practically, it means that our identification problem may lead to drifting behavior in the sampler if identifying restrictions upon utility are not given.

We assume a cardinal utility framework. Specifically, we can make the additional assumption that $F(.)$ is known up to a linear transformation:

$$F(U) = a + bU \tag{2.35}$$

Although we cannot identify the translation and scale parameter, we can impose restrictions like $F(0) = 0$ and $\frac{\partial F(U)}{\partial U} = 1$ for all U to identify our transformation.

The restriction $F(0) = 0$ (i.e. $a = 0$) is achieved through the untestable condition of weak complementarity (Maler, 1974), in which consumers receive no utility from a non-essential good's attributes if they do not consume it. In other words, we impose $U_j(0) = 0$ for all inside goods $j = 1, ..., J$ and $U_z(0) = 0$ for the outside good. C. R. Bhat (2008) also makes this assumption and explains that it essentially represents a cardinal normalization restriction on utilities, and since a cardinal restriction on preferences must be eventually used for welfare measurement, using weak complementarity makes sense here.

The restriction $F'(U) = 1$ is achieved by rewriting $\frac{\partial F}{\partial U}$. Let the $n+1, n+2, ..., J^{th}$ inside goods, and the outside good z being consumed for a given purchase occasion.

Then we have:

$$\log\left(\frac{\partial U_j(q_{jt})}{\partial q_j}\right) + \log\left(\frac{\partial F}{\partial U}\right) = \log(\lambda) + \log(p_j) \text{ if } j \text{ s.t. } q_j > 0 \tag{2.36}$$

$$\log\left(\frac{\partial U_j(q_{jt})}{\partial q_j}\right) + \log\left(\frac{\partial F}{\partial U}\right) < \log(\lambda) + \log(p_j) \text{ if } j \text{ s.t. } q_j = 0 \tag{2.37}$$

which implies, assuming that the first good is always consumed without loss of generality, and after differencing:

$$\log\left(\frac{\partial U_j(q_{jt})}{\partial q_j}\right) - \log\left(\frac{\partial U_j(q_{1t})}{\partial q_1}\right) = \log(p_j) - \log(p_1) \text{ if } j \text{ s.t. } q_j > 0 \tag{2.38}$$

$$\log\left(\frac{\partial U_j(q_{jt})}{\partial q_j}\right) - \log\left(\frac{\partial U_j(q_{1t})}{\partial q_1}\right) < \log(p_j) - \log(p_1) \text{ if } j \text{ s.t. } q_j = 0 \tag{2.39}$$

for all $j = 2, \ldots, J$. Fixing the scaling of the log-marginal utility for one of the goods allows us to removes the non-identification associated with $\log\left(\frac{\partial F}{\partial U}\right)$. This is achieved in the simulation exercises below by using a known functional form for the inside goods, and setting the baseline utility parameter for one of the goods to 1 without loss of generality. A parametric utility for inside goods is necessary when the number of distinct price points and quantity points is small, as it is often the case in practice. In the theoretical case where the inside goods' preferences are strongly informed by the data – large number of distinct price points and quantities purchased – then the marginal utility of one of the inside goods must be set by the analyst. The functional form for the sub-utilities are identified by variation in purchase shares for each good, and the contemporaneous variation in purchase quantities.

2.3.5 Simulation Exercises

We build two simulation studies to recover the latent preferences of an individual consumer. In the first study, we assume that the analyst knows the correct functional form of the outside good, and estimates two models: the baseline parametric model, and the model with Gaussian process prior on the outside good. In the second study, we assume that the analyst does not know the correct functional form of the outside good utility function, and estimates two models: the baseline parametric model, and the model with Gaussian process prior on the outside good. In both studies and both models, the inside good utility function is assumed to be correctly known: we use the following sub-utility for each inside good $j = 1, \ldots, J$:

$$\nu_j(q_j) = \frac{\psi_j}{\gamma_j}\log(\gamma_j q_j + 1) \tag{2.40}$$

where J is the number of inside goods. The parameter γ introduces flexibility in the satiation rate (C. R. Bhat, 2005). Following the above-mentioned data generating process, we use synthetic data for two inside goods and one outside good. Prices are uniformly drawn between 0.2 and 5.0, and the price of the outside good is fixed to 1 (numeraire). The total consumer budget is also drawn at each time period from a uniform distribution between \$1.0 and \$20. This stochastic budget illustrates the non-stationarity of the total budget consumers allocates at each purchase occasion. The data is generated via an interior-point optimizer, suited for large-scale nonlinear optimization. We consider the case of one consumer with fixed preferences in 900 purchase occasions, such that we observe as many vectors of prices and quantities. The standard deviation of the error term is assumed to be 0.1.

We estimate this model with the no-U-turn sampler (NUTS) variant of Hamiltonian Monte Carlo (Hoffman and Gelman, 2014), sampling simultaneously the parameters from the inside goods' subutility functions, and the latent functions from the Gaussian process in the outside good's subutility function. The MCMC algorithm is run for 2,000 iterations including 1,000 burn-in iterations, using 10 independent chains. We make predictions on the range of the data. The model is able to recover each subutility function and provides us with 95% credible intervals, which are constructed from empirical 2.5^{th} and 97.5^{th} percentiles.

Comparing Parametric and GP Prior With Correct Functional Form

In this study, the sub-utility associated with the outside good is assumed to be

$$u_z(z) = \psi_3 \log(z) \tag{2.41}$$

such that the quantity consumed is always strictly positive. Table 2.1 shows the posterior summary of parameters for correctly specified parametric prior placed on the outside good's marginal utility. Table 2.2 shows the posterior summary of parameters for Gaussian prior placed on the outside good's marginal utility and second derivative of utility. Figure 2.1 explains how the parametric prior compares to the Gaussian process prior on marginal utility for the same simulated observations. Results are comparable although the Gaussian process model exhibits higher uncertainty in the functional form, since it is being estimated from the data. Table 2.2 and Figure 2.1 act as a sanity check.

Comparing Parametric and GP Prior With Misspecified Functional Form: Slower Satiation

In this study, the sub-utility associated with the outside good is assumed to be

$$u_z(z) = \psi_3(\log(z) + z) \tag{2.42}$$

such that the quantity consumed is always strictly positive. However, this time, utility satiates slower, as marginal utility is higher by an additional ψ_3. Table 2.3 shows

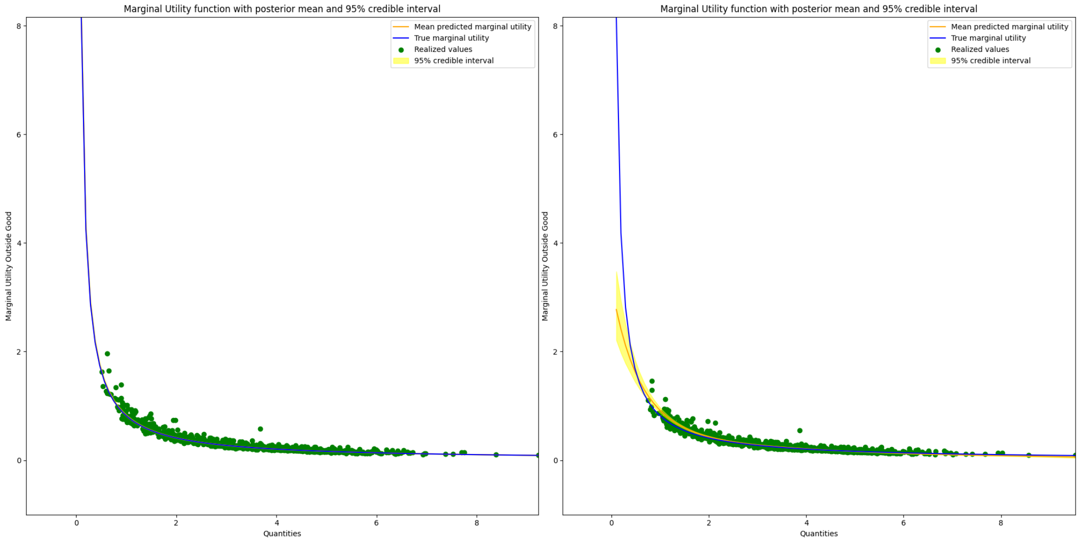
Parameter	Mean	SD	HDI 3%	HDI 97%	MCSE (Mean)	MCSE (SD)	ESS (Bulk)	ESS (Tail)	$\hat{R}$
ψ_1	1.000	0.000	1.000	1.000	0.0	0.0	10000.0	10000.0	-
ψ_2	1.739	0.029	1.685	1.793	0.0	0.0	3590.0	4260.0	1.0
ψ_3	0.823	0.009	0.807	0.840	0.0	0.0	3522.0	4210.0	1.0
γ_1	0.975	0.019	0.940	1.011	0.0	0.0	3419.0	3938.0	1.0
γ_2	0.990	0.020	0.951	1.029	0.0	0.0	4357.0	4564.0	1.0

TABLE 2.1: Posterior Summary of Parameters for correctly specified parametric prior placed on outside good's marginal utility. Two inside goods and one outside good are used. ψ_1 is correctly set to 1 for identification. The ground truth for ψ is $[1, 1.74, 0.81]$ and for γ is $[1, 1]$. Mean: posterior mean; SD: posterior standard deviation; HDI: high density posterior interval; MCSE: Monte Carlo standard error; ESS: effective sample size; $\hat{R}$: R-hat statistic.

Parameter	Mean	SD	HDI 3%	HDI 97%	MCSE (Mean)	MCSE (SD)	ESS (Bulk)	ESS (Tail)	$\hat{R}$
ψ_1	1.000	0.000	1.000	1.000	0.000	0.000	10000.0	10000.0	-
ψ_2	1.723	0.040	1.647	1.795	0.001	0.001	2735.0	3518.0	1.00
γ_1	0.975	0.043	0.893	1.055	0.001	0.000	4160.0	5500.0	1.00
γ_2	0.978	0.029	0.924	1.033	0.000	0.000	4340.0	6276.0	1.00
intercept	1.688	1.720	-1.145	5.592	0.062	0.058	1094.0	342.0	1.01
slope	-0.076	0.066	-0.199	-0.000	0.002	0.001	1628.0	2089.0	1.00
σ_f	4.678	4.510	0.021	12.947	0.108	0.076	1443.0	2719.0	1.00
β	8.473	2.219	4.535	12.374	0.061	0.043	1342.0	2538.0	1.01

TABLE 2.2: Posterior summary of parameters for Gaussian process prior placed on outside good's marginal utility. Two inside goods and one outside good are used. The ground truth for ψ is $[1, 1.74, 0.81]$ and for γ is $[1, 1]$. ψ_1 is correctly set to 1 for identification. The intercept and slope parameters of the Gaussian process correspond to an affine functional form for the mean of the Gaussian process. σ_f is the square root of the signal variance hyperparameter, and β is the lengthscale hyperparameter, which are estimated. Mean: posterior mean; SD: posterior standard deviation; HDI: high density posterior interval; MCSE: Monte Carlo standard error; ESS: effective sample size; $\hat{R}$: R-hat statistic.

the posterior summary of parameters for misspecified parametric prior placed on the outside good's marginal utility. Table 2.4 shows the posterior summary of parameters for Gaussian prior placed on the outside good's marginal utility and second derivative of utility. Here we observe that the Gaussian process specification is robust to misspecification of the functional form of the outside good's utility. Figure 2.2 shows the inability of the parametric model to correctly fit the misspecified marginal utility, as opposed to the Gaussian process prior that is enable to capture the slower satiation rate.



(a) Parametric

(b) Gaussian process

FIGURE 2.1: Estimated marginal utility of the outside good when the correctly parametric functional form $u_z(z) = \psi \log(z)$ (i.e., $u'_z(z) = \psi/z$) is imposed (a) and when a Gaussian process prior is used on the latent utility function (b). The Gaussian process specification performs comparatively to the correctly specified parametric prior, but exhibits higher uncertainty due to an estimation more demanding of the data. Note that the realized utility values are unobserved by the analyst.

2.4 Application 2: Nonhomothetic Discrete Choice Model with Household Production

2.4.1 Overcoming Separability with Household Production Theory

Separability and additive preferences are core concepts in consumer theory that significantly impact the modeling of consumer choice behavior. Separability refers to the idea that the utility derived from consuming a group of goods is independent of the consumption of other goods. In other words, the consumer's preference for one set of goods does not influence their preference for another set. This notion extends to additive preferences, where the overall utility a consumer derives from consuming all goods is simply the sum of the utilities derived from each good independently. This framework simplifies the analysis by allowing economists to consider each good or group of goods in isolation, without needing to account for complex interactions between different consumption choices.

However, the assumption of separability poses significant challenges when modeling consumer behavior across multiple product categories. Real-world observations frequently demonstrate that consumers' decisions about one category can be influenced by their choices in another, indicating a level of interdependence that separability cannot capture. For example, the purchase of a high-end coffee maker

Parameter	Mean	SD	HDI 3%	HDI 97%	MCSE (Mean)	MCSE (SD)	ESS (Bulk)	ESS (Tail)	$\hat{R}$
ψ_1	1.000	0.000	1.000	1.000	0.000	0.000	10000.0	10000.0	-
ψ_2	1.625	0.045	1.544	1.712	0.001	0.001	3603.0	4471.0	1.0
ψ_3	9.722	0.235	9.278	10.153	0.004	0.003	3902.0	4741.0	1.0
γ_0	0.811	0.052	0.717	0.912	0.001	0.001	4409.0	5505.0	1.0
γ_1	0.709	0.029	0.654	0.764	0.000	0.000	5112.0	5551.0	1.0

TABLE 2.3: Posterior summary of parameters for misspecified parametric prior placed on outside good's marginal utility. Two inside goods and one outside good are used. ψ_1 is correctly set to 1 for identification. The ground truth for ψ is $[1, 1.74, 0.81]$ and for γ is $[1, 1]$. Mean: posterior mean; SD: posterior standard deviation; HDI: high density posterior interval; MCSE: Monte Carlo standard error; ESS: effective sample size; $\hat{R}$: R-hat statistic.

might increase a consumer's preference for premium coffee beans, suggesting a complementarity that separable models overlook. This interdependence between product categories means that consumer preferences are, in fact, nonseparable, making it difficult to accurately predict consumer behavior using models that rely on the assumption of separability. As a result, the use of separable models can lead to incorrect inferences on demand, consumer welfare, and suboptimal marketing strategies.

The Household Production Theory offers a compelling framework to address these limitations. Originally proposed by Becker (1965), this theory posits that households derive utility not directly from market goods, but from "commodities" they produce using these goods as inputs, along with their time – although, for simplicity, we do not consider time as input, in this chapter. The theory provides a broader perspective on consumer choice, emphasizing the process by which households combine purchased goods with their labor to produce final commodities that directly contribute to utility. This approach has been further elaborated in the literature, with scholars such as Pollak and Wachter (1975) and Gronau (1977) exploring its implications for labor supply, time allocation, and the valuation of non-market activities. By focusing on the production process within the household, this theory introduces a natural framework for considering the interactions between different product categories and the nonseparability of preferences in a more nuanced manner.

In this application, we propose a model that, while initially based on separable preferences, is enhanced by incorporating intermediary goods, thus accommodating nonseparability in consumer preferences across multiple product categories. Our approach leverages the Household Production Theory to understand how the consumption of intermediary goods—those used in the production of final commodities within the household—can create linkages between seemingly independent product categories. By integrating these goods into our model, we demonstrate how preferences for one category of products can influence preferences for another, overcoming the limitations imposed by the traditional assumption of separability. This novel

Parameter	Mean	SD	HDI 3%	HDI 97%	MCSE (Mean)	MCSE (SD)	ESS (Bulk)	ESS (Tail)	$\hat{R}$
ψ_1	1.000	0.000	1.000	1.000	0.000	0.000	10000.0	10000.0	-
ψ_2	1.723	0.040	1.646	1.798	0.001	0.001	3224.0	5195.0	1.00
γ_1	0.974	0.044	0.893	1.057	0.001	0.001	3731.0	5013.0	1.00
γ_2	0.978	0.029	0.924	1.031	0.000	0.000	4505.0	5629.0	1.00
intercept	1.667	1.664	-1.217	5.036	0.047	0.033	1332.0	2332.0	1.01
slope	-0.075	0.066	-0.197	-0.000	0.001	0.001	2058.0	2359.0	1.01
σ_f	4.485	4.451	0.013	13.014	0.102	0.072	1351.0	2176.0	1.00
β	8.394	2.302	4.352	12.464	0.069	0.049	1065.0	1627.0	1.01

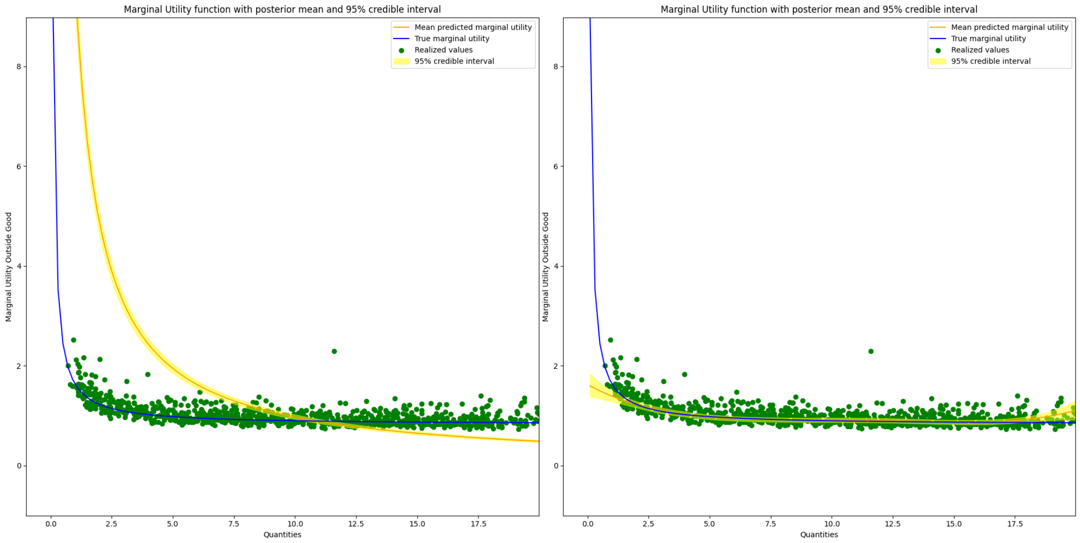
TABLE 2.4: Posterior summary of parameters for Gaussian process prior placed on outside good's marginal utility. Two inside goods and one outside good are used. The ground truth for ψ is $[1, 1.74, 0.81]$ and for γ is $[1, 1]$. ψ_1 correctly is set to 1 for identification. The ground truth for the outside good is a utility function of the following form: $u_z(z) = \psi(\log(z) + z)$. The intercept and slope parameters of the Gaussian process correspond to an affine functional form for the mean of the Gaussian process. σ_f is the square root of the signal variance hyperparameter, and β is the lengthscale hyperparameter, which are estimated. Mean: posterior mean; SD: posterior standard deviation; HDI: high density posterior interval; MCSE: Monte Carlo standard error; ESS: effective sample size; $\hat{R}$: R-hat statistic.

contribution not only aligns our model more closely with observed consumer behavior but also offers new insights into the complex interplay between different types of goods in household production and consumption processes.

We need to better understand how complementarity between different consumer packaged goods is moderated by expansion in the product category. Take the example of burgers and buns, which function as both intermediary and final goods. When they combine to form a sandwich, the final good – demand is expected to rise with consumer expenditure, highlighting as complex complementarity influenced by expenditure.

The role of outside good consumption further complicates choice modeling regarding budget allocation and price effects. Traditional models, imposing linear utility from outside goods, assume that consumers do not satiate in their outside good consumption. This unrealistic assumption fails to account for diminishing returns of utility in consumption – a scenario better captured by nonlinear utility models. However, even nonlinear utility models make parametric assumptions that cannot be justified *ex ante*.

In this section, we relax a parametric assumption on the outside good consumption, allowing the detection of more flexible income effects. We formulate a structural model of household production and consumption, using constrained utility maximization. The consumer problem is to buy the optimal volume of intermediary



(a) Parametric

(b) Gaussian process

FIGURE 2.2: Estimated marginal utility of the outside good when the misspecified parametric functional form $u_z(z) = \psi \log(z)$ (i.e., $u_z'(z) = \psi/z$ is imposed (a) and when a Gaussian process prior is used on the latent utility function (b). The ground truth is a marginal utility function of the following form: $u_z(z) = \psi(\log(z) + z)$ (i.e., $u_z'(z) = \psi/z + \psi$). The Gaussian process specification is robust to an unknown functional form of outside good preference. Note that the realized utility values are unobserved by the analyst.

goods $\mathbf{q}$ and outside good z, and use the intermediary goods to produce a utility-maximizing final good to consume.

$$
\begin{aligned}
V(\mathbf{p}, E) = \max_{\mathbf{c},\mathbf{q},\mathbf{z}} u(\mathbf{c}, z) &= \ln u(\mathbf{c}) + \tau \ln(z) \\
&\text{subject to:} \\
&\quad \mathbf{p}' q + z \leq E \\
&\quad \mathbf{q} \in \mathcal{Q} \\
&\quad \mathbf{c} \in \mathcal{C}(\mathbf{q})
\end{aligned}
\tag{2.43}
$$

where $\mathbf{c} \in \mathcal{C}(\mathbf{q})$ is the set of final goods that consumers are able to produce, $\mathbf{p}$ is the price vector, and E is the budget allotment and the price of the outside good is assumed to be \$1.00 without loss of generality. The production-consumption step allows us to parameterize $U(c, z)$ and $\mathcal{C}(\mathbf{q})$ instead of utility in terms of quantities purchased.

Consumers always purchase some of the outside good, but purchase exactly one product from the final good category. Utility maximization results in a corner solution; therefore, a linear utility function is appropriate:

$$
u(\mathbf{c}) = \boldsymbol{\psi}' \mathbf{c} \tag{2.44}
$$

Because we are interested in understanding how consumers switch between different final consumption goods as E increases, we need utility to be *nonhomothetic*. Consumers obtain higher utility by changing their demand toward a higher-quality final good, instead of buying more units of the same lesser-quality final good.

The utility for the vector of demand of the final good is defined implicitly as:

$$u(\mathbf{c}) = \sum_{k=1}^{K} \psi_k(\bar{u}) c_k = \sum_{k=1}^{K} \exp\left(\alpha_k - \kappa_k \bar{u}(\mathbf{c}, z)\right) c_k \tag{2.45}$$

where the marginal utility of a final good is a function of attainable utility $\bar{u}$. Allenby and Rossi (1991) and Allenby, Garratt, and Rossi (2010) use a similar utility model as it allows for superior good effect and account for expenditure effects by rotating the indifference curves as utility increases with expenditure.

Our parameterization of utility implies the following utility for purchased goods. The consumer problem is then rewritten as follows:

$$\begin{aligned} &\max_{\mathbf{c},z} u(\mathbf{c}, z) = \ln u(\mathbf{c}) + \tau \ln(z) \\ &\text{s.t.} \quad \sum_{k=1}^{K} f_k c_k + z \leq E \end{aligned} \tag{2.46}$$

The input-output matrix A considers intermediary goods quantities a_{jk} such that the volume of input j is required to make one unit of final good k. As an example, we consider the case where $J = 2$ and $K = 3$, where intermediary goods are buns and burgers, and final goods are bun, burger, and sandwich:

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \tag{2.47}$$

In that case, there are three possible final goods, c_1 (bun), c_2 (burger) and c_3 (sandwich). The set $\mathcal{C}(\mathbf{q})$ of final good quantities that can be produced from input volumes $\mathbf{q}$ is the set of vector $\mathbf{c}$ with nonnegative entries and such that $\sum_k a_{jk} c_k \leq q_j$ for all j.

Under the assumption that consumers do not keep inventories, buying more intermediary goods only come at an extra cost; therefore the demand for intermediary goods be such that:

$$q_j^* = \sum_{k=1}^{K} a_{jk} c_k^* = a_{jk} \mathbb{1}(c_k^* = 1). \tag{2.48}$$

for all $j = 1, \ldots, J$. We derive the full price f_k of each final good k as the dollar amount that consumers need to pay to produce a unit of that final good:

$$f_k = \sum_{j=1}^{J} a_{jk} p_j \tag{2.49}$$

Likelihood specification

Adding a multiplicative error to equation (2.45), we obtain the probability of selecting final good k:

$$\Pr(c_k = 1) = \Pr(\alpha_k - \kappa_k \bar{u}^k + \tau \ln(E - f_k) + \varepsilon_k > \alpha_i - \kappa_i \bar{u}^i + \tau \ln(E - f_i) + \varepsilon_i) \tag{2.50}$$

for all i such that $p_i \leq E$. Assuming type 1 extreme value errors with scale parameter 1 (i.e., variance $\frac{\pi^2}{6}$):

$$\Pr(c_k = 1) = \frac{\exp(\alpha_k - \kappa_k \bar{u}^k + \tau \ln(E - f_k))}{\sum_{i|f_i<E} \exp(\alpha_i - \kappa_i \bar{u}^i + \tau \ln(E - f_i))} \tag{2.51}$$

which we can rewrite, by changing the parametrization to intermediary goods:

$$\Pr(q_j = a_{jk}) = \frac{\exp(\alpha_k - \kappa_k \bar{u}^k + \tau \ln(E - \sum_{j=1}^{J} a_{jk} p_j))}{\sum_{i|\sum_{j=1}^{J} a_{ji} p_j < E} \exp(\alpha_i - \kappa_i \bar{u}^i + \tau \ln(E - \sum_{j=1}^{J} a_{ji} p_j))} \tag{2.52}$$

Nonparametric specification

We relax the functional form on the outside good to

$$\begin{aligned} V(\mathbf{p}, E) = \max_{\mathbf{c},\mathbf{q},\mathbf{z}} u(\mathbf{c}, z) &= \ln u(\mathbf{c}) + u_z(z) \\ \text{subject to:}& \\ \mathbf{p}'q + z &\leq E \\ \mathbf{q} &\in \mathcal{Q} \\ \mathbf{c} &\in \mathcal{C}(\mathbf{q}) \end{aligned} \tag{2.53}$$

where $u_z(z)$ is an unknown function of the outside good whose functional form is estimated from the data. We place a Gaussian process (GP) prior on $u_z(.)$:

$$u_z(.) \sim \mathcal{GP}(\boldsymbol{\mu}, \mathbf{K}) \tag{2.54}$$

where we have the mean and covariance function defined as follows:

$$\boldsymbol{\mu}(\mathbf{z}) = \mathbb{E}\left[u(\mathbf{z})\right] \tag{2.55}$$

$$\mathbf{K}(\mathbf{z}, \mathbf{z}') = \mathbb{E}\left[\left(u(\mathbf{z}) - \boldsymbol{\mu}(\mathbf{z})\right)\left(u(\mathbf{z}') - \boldsymbol{\mu}(\mathbf{z}')\right)\right] \tag{2.56}$$

Notice our GP looks like a regression model in function-space view (see Rasmussen, Williams, et al. (2006), §2.3). We use the squared exponential (SE) covariance function for the kernel:

$$K(z_t, z_{t'}) = \sigma_f^2 \exp\left(-\frac{1}{2\beta^2}(z_t - z_{t'})^2\right) \tag{2.57}$$

where β denotes the characteristic length-scale and σ_f^2 is the signal variance, hyperparameters of the GP model. The squared exponential covariance function implies that the covariance is almost unity (when $\sigma_f^2 = 1$) between variables whose corresponding inputs are very close (Rasmussen, Williams, et al., 2006)[2] Adding a multiplicative error to equation (2.45), we obtain the probability of selecting final good k:

$$\Pr(c_k = 1) = \Pr(\alpha_k - \kappa_k \bar{u}^k + u_z(E - f_k) + \varepsilon_k > \alpha_i - \kappa_i \bar{u}^i + u_z(E - f_i) + \varepsilon_i) \quad (2.58)$$

Assuming type 1 extreme value errors with scale parameter 1 (i.e., variance $\frac{\pi^2}{6}$):

$$\Pr(c_k = 1) = \frac{\exp(\alpha_k - \kappa_k \bar{u}^k + u_z(E - f_k))}{\sum_{i|f_i<E} \exp(\alpha_i - \kappa_i \bar{u}^i + u_z(E - f_i))} \quad (2.59)$$

which we can rewrite, by changing the parametrization to intermediary goods:

$$\Pr(q_j = a_{jk}) = \frac{\exp(\alpha_k - \kappa_k \bar{u}^k + u_z(E - \sum_{j=1}^J a_{jk} p_j))}{\sum_{i|\sum_{j=1}^J a_{ji} p_j < E} \exp(\alpha_i - \kappa_i \bar{u}^i + u_z(E - \sum_{j=1}^J a_{ji} p_j))} \quad (2.60)$$

We derive below the derivative of demand and show that own and cross-price effects all depend on the marginal utility of the outside good. Hence, misspecification of that marginal utility leads to inconsistent estimation of own- and cross-price effects. We also show that using intermediary goods enable non-trivial own- and cross-price effects, even when using an additive utility framework for final goods.

Derivatives of Demand

$$\frac{\partial \Pr(c_k = 1)}{\partial f_l} = \begin{cases} -\left(I(k=l)\Pr(c_k = 1) - \Pr(c_k = 1)\Pr(c_l = 1)u_z'(E - f_l)\right)\left(\frac{1}{1+\kappa_l u^l}\right) \\ 0 \ \text{if} f_l > E \end{cases} \quad (2.61)$$

Then, we calculate the derivative of the intermediary demand with respect to prices. We first consider the case where the good m and good k are used to make

[2] It has been shown that the squared exponential covariance function corresponds to a Bayesian linear regression model with an infinite number of basis function (*ibid*), which emphasizes the flexibility of such a function. Another advantage of this covariance function is that it is infinitely differentiable, which is useful in our context. Finally, we also choose this function for the interpretability of its hyperparameters. The characteristics length-scale parameter β controls the amount of information that the Gaussian process will borrow around the test input point that needs to be evaluated. A large β means that more information is borrowed, which will smooth the utility function; conversely, a smaller β will make the utility function more prone to capture non-linearities in preferences and rationalize them. The characteristic length-scale β and the signal variance σ_f^2 are weakly identified and their proportion is more important to the predictive performance than their individual value for the Matérn class of covariance function (Diggle, Tawn, and Moyeed, 1998; H. Zhang, 2004), to which belongs the squared exponential covariance function.

final good k: We can re-parameterize this derivative:

$$\begin{aligned}
\frac{\partial \Pr(q_j = a_{jk})}{\partial p_{j'}} &= \sum_{k'=1}^{K} \frac{\partial \Pr(q_j = a_{jk})}{\partial f_{k'}} \frac{\partial f_{k'}}{\partial p_{j'}} && (2.62)\\
&= \sum_{k'=1}^{K} \frac{\partial \Pr(c_k = 1)}{\partial f_{k'}} \frac{\partial \sum_{j=1}^{J} a_{jk'} p_j}{\partial p_{k'}} && (2.63)\\
&= \sum_{k'=1}^{K} \frac{\partial \Pr(c_k = 1)}{\partial f_{k'}} a_{jk'} && (2.64)\\
&= \frac{\partial \Pr(c_k = 1)}{\partial f_k} a_{jk} + \sum_{\substack{k'=1\\ k'\neq k}}^{K} \frac{\partial \Pr(c_k = 1)}{\partial f_{k'}} a_{jk'} && (2.65)
\end{aligned}$$

The expected optimal demand for intermediary good j is:

$$\begin{aligned}
\mathbf{E}(q_j^*) &= \mathbf{E}\left(\sum_{k=1}^{K} a_{jk} c_k^*\right) && (2.66)\\
&= \mathbf{E}\left(\sum_{k=1}^{K} a_{jk} c_k^*\right) && (2.67)\\
&= \sum_{k=1}^{K} a_{jk} \Pr(c_k^* = 1) && (2.68)
\end{aligned}$$

We can compute the derivatives of expected demand with respect to prices:

$$\begin{aligned}
\frac{\partial \mathbf{E}(q_j^*)}{\partial p_l} &= \sum_{k=1}^{K} a_{jk} a_{lk} \frac{\partial \Pr(c_k = 1)}{\partial f_k} + \sum_{k=1}^{K} \sum_{\substack{k'=1\\ k\neq k}}^{K} a_{jk} a_{lk'} \frac{\partial \Pr(c_k = 1)}{\partial f_{k'}} && (2.69)\\
&= -\sum_{k=1}^{K} a_{jk} a_{lk} \frac{1}{1+\kappa_k u_k} u_z'(E - f_k) \Pr(c_k = 1)(1 - \Pr(c_k = 1))\\
&+ \sum_{k=1}^{K} \sum_{\substack{k'=1\\ k\neq k}}^{K} a_{jk} a_{lk'} \frac{1}{1+\kappa_{k'} u_{k'}} u_z'(E - f_{k'}) \Pr(c_k = 1) \Pr(c_k' = 1) && (2.70)
\end{aligned}$$

In the case where $l = j$, the effect is non-obvious, except when $a_{jk'} = 0$ for all $k' \neq k$, that is, without loss of generality, when the intermediary good j is used to produce exclusively the final good k and not any other final good; in that case, demand of j decreases with price of good j. Conversely, it is possible that demand of j increases with price of good j, provided that the first element in the sum in (2.70) is smaller, in absolute terms, than the second element. This case may happen, e.g., when the intermediary good j is used to make simultaneously many final goods; these competing final goods becoming relatively more attractive, and that the demand of other intermediary goods becomes so high that it calls for more intermediary good j by complementarity effect.

We estimate this model with the no-U-turn sampler (NUTS) variant of Hamiltonian Monte Carlo (Hoffman and Gelman, 2014), sampling simultaneously the parameters from the inside goods' subutility functions, and the latent functions from the Gaussian process in the outside good's subutility function. The implicit utility values in the likelihood are obtained using Newton's method at each MCMC iteration (Allenby, Garratt, and Rossi, 2010).

2.4.2 Identification

The outside good utility functional form is identified nonparametrically by variation in purchase shares for each inside good. However, the outside good utility can be translated by a factor a and rotated by a scale parameter b without changing the likelihood function, which causes an identification issue. Let $v_z(.) \equiv a + bu_z(.)$. Assuming a type 1 extreme value error distribution with scale σ:

$$\Pr(q_j = a_{jk}) = \frac{\exp\left(\frac{1}{\sigma}(\alpha_k - \kappa_k \bar{u}^k + v_z(E - \sum_{j=1}^J a_{jk} p_j))\right)}{\sum_{i|\sum_{j=1}^J a_{ji} p_j < E} \exp\left(\frac{1}{\sigma}(\alpha_i - \kappa_i \bar{u}^i + v_z(E - \sum_{j=1}^J a_{ji} p_j))\right)} \tag{2.71}$$

$$= \frac{\exp\left(\frac{1}{\sigma}(\alpha_k - \kappa_k \bar{u}^k + a + bu_z(E - \sum_{j=1}^J a_{jk} p_j))\right)}{\sum_{i|\sum_{j=1}^J a_{ji} p_j < E} \exp\left(\frac{1}{\sigma}(\alpha_i - \kappa_i \bar{u}^i + a + bu_z(E - \sum_{j=1}^J a_{ji} p_j) + b)\right)} \tag{2.72}$$

$$= \frac{\exp\left(\frac{1}{\sigma'}(\alpha_k - \kappa_k \bar{u}^k + u_z(E - \sum_{j=1}^J a_{jk} p_j))\right)}{\sum_{i|\sum_{j=1}^J a_{ji} p_j < E} \exp\left(\frac{1}{\sigma'}(\alpha_i - \kappa_i \bar{u}^i + u_z(E - \sum_{j=1}^J a_{ji} p_j))\right)} \tag{2.73}$$

where $\sigma' = \sigma / b$. In theory, a is identified by setting the scale parameter of the error term to 1, but in practice, in our experiments, a is not always well identified by the data only. As a consequence, a and b need to be fixed by the analyst. A convenient way to fix these values is to set the slope and the intercept by conditioning on two observations *in the prior distribution* evaluated at set, assumed known utility values, without loss of generality. The joint distribution of the training vector of utility and the vector of utility at $\boldsymbol{z}_*$ is:

$$\begin{bmatrix} u_z(\boldsymbol{z}) \\ u_z(\boldsymbol{z}_*) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}(\boldsymbol{z}) \\ \boldsymbol{\mu}(\boldsymbol{z}_*) \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}(\boldsymbol{z}, \boldsymbol{z}) & \boldsymbol{K}(\boldsymbol{z}, \boldsymbol{z}_*) \\ \boldsymbol{K}(\boldsymbol{z}_*, \boldsymbol{z}) & \boldsymbol{K}(\boldsymbol{z}_*, \boldsymbol{z}_*) \end{bmatrix}\right) \tag{2.74}$$

Let $\boldsymbol{z}$ be the training data that is being observed. We condition on two observations $\boldsymbol{z}_*$ by conditioning the joint Gaussian prior distributions on the observations to give:

$$u_z(\boldsymbol{z}) \mid u_z(\boldsymbol{z}_*) \sim \mathcal{N}\left(\boldsymbol{\mu}', \mathbf{K}'\right) \tag{2.75}$$

where

$$\mu' = \boldsymbol{\mu}(\boldsymbol{z}) + \mathbf{K}(\boldsymbol{z}_*, \boldsymbol{z})\mathbf{K}(\boldsymbol{z}_*, \boldsymbol{z}_*)^{-1}(u_z(\boldsymbol{z}_*) - \boldsymbol{\mu}(\boldsymbol{z}_*)) \tag{2.76}$$

and

$$\mathbf{K}' = \mathbf{K}(\boldsymbol{z}, \boldsymbol{z}) - \mathbf{K}(\boldsymbol{z}_*, \boldsymbol{z})\mathbf{K}(\boldsymbol{z}_*, \boldsymbol{z}_*)^{-1}\mathbf{K}(\boldsymbol{z}, \boldsymbol{z}_*) \tag{2.77}$$

2.4.3 Simulation Exercises

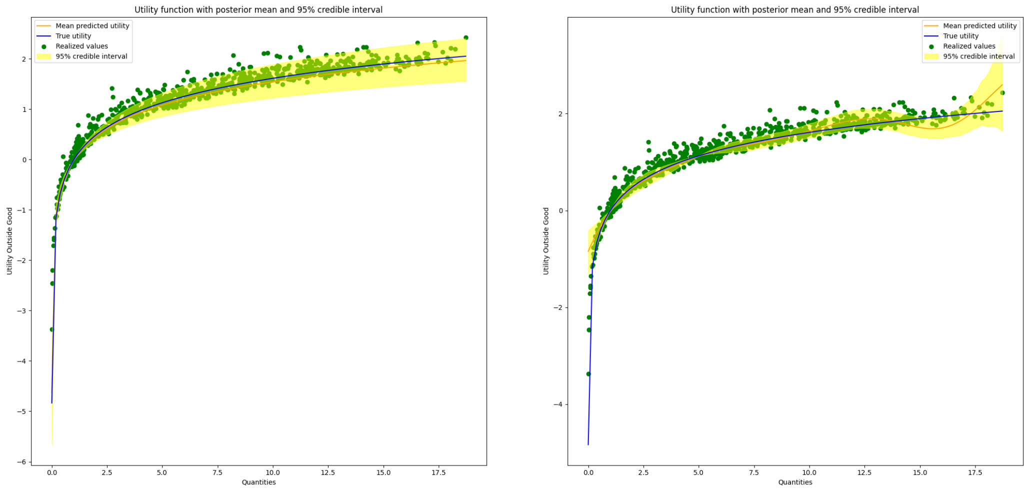
We build four simulation studies to recover the latent preferences of an individual consumer. In the first study, we assume that the analyst knows the correct functional form of the outside good, and estimates two models: the baseline parametric model, and the model with Gaussian process prior on the outside good. In the subsequent studies, we assume that the analyst does not know the correct functional form of the outside good utility function, and estimates two models: the baseline parametric model, and the model with Gaussian process prior on the outside good.

Following the above-mentioned data generating process, we use synthetic data for three final goods (e.g., burgers, buns, and sandwiches) Prices are uniformly drawn between 1.0 and 5.0, and the price of the outside good is fixed to 1 (numeraire). The total consumer budget is also drawn at each time period from a uniform distribution between \$1.0 and \$5, but correcting *ex post* to make sure that each allocation is feasible, i.e., at least one inside good is purchased. We consider the case of one consumer with fixed preferences in 900 training purchase occasions, such that we observe as many vectors of prices and quantities. We also set aside 100 testing purchase occasions to understand how each specification generalizes. The scale of the Type 1 extreme value error term is assumed to be small (0.1).

We estimate this model with the no-U-turn sampler (NUTS) variant of Hamiltonian Monte Carlo (Hoffman and Gelman, 2014), sampling simultaneously the parameters from the inside goods' subutility functions, and the latent functions from the Gaussian process in the outside good's subutility function. The MCMC algorithm is run for 1,100 iterations including 1,000 burn-in iterations, using 24 independent chains. We make predictions on the range of the data. The model is able to recover each subutility function and provides us with 95% credible intervals, which are constructed from empirical 2.5^{th} and 97.5^{th} percentiles.

Figure 2.3 acts as a sanity check and shows that the Gaussian process model is on par with the correctly specified parametric model, especially where the observations are more dense. We note that when the data gets scare (around 15 units), the Gaussian process becomes more unstable and tends to deviate from the ground truth (in blue and green).

In a second study, we use a more exotic functional form of outside good utility $u(z) = \sin(0.7z) + 0.7z$ which plateaus at specific regions of the space (Figure 2.4). We observe that the misspecified parametric prior struggles to capture the nonlinear rate of satiation, as opposed to the Gaussian process prior. It's important to realize that estimation proceeds from the vector of price, total expenditure, and quantities (zero or one) from the final goods. The data requirements are minimal to estimate each model. The nonlinear rate of satiation is estimated from the sudden variations



(a) Parametric

(b) Gaussian process

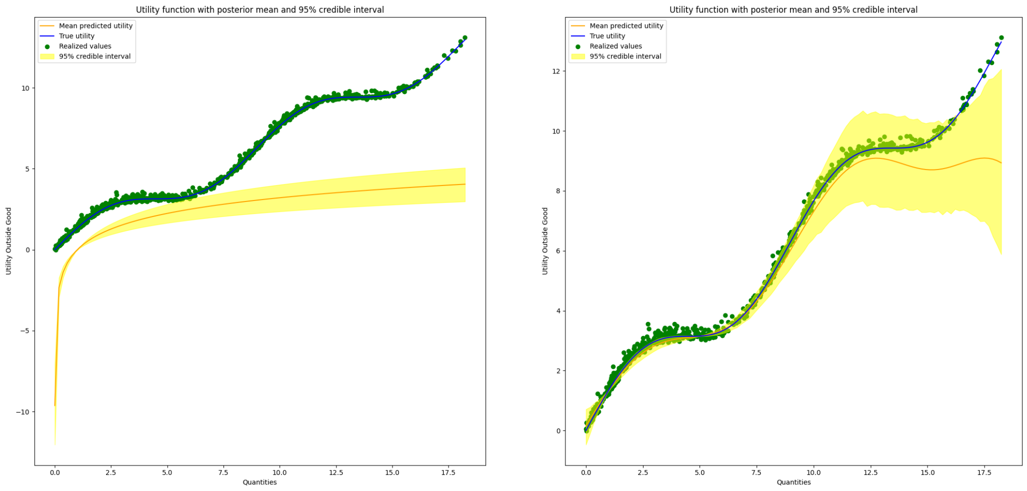
FIGURE 2.3: Estimated utility of the outside good when the correct parametric functional form is imposed (a) and when a Gaussian process prior is used on the latent utility function (b). The ground truth is a utility function of the power form: $u_z(z) = z^{0.7}$. Note that the realized utility values are unobserved by the analyst. The hyperparameters of the Gaussian process are manually set and not estimated, for simplicity and as it significantly reduces the computational burden.

in probabilities of chosen goods when expenditure increases, characterized by different empirical choice distributions with increased expenditure.

In a third study, we use an outside good functional form $u(z) = 3 - \exp(2 - 0.7z)$, which displays less satiation at first, and more aggressive satiation after a threshold is reached (Figure 2.5). Once more, the parametric functional form struggles to fit the change in satiation rate and compromises across the two rates. On the other hand, the Gaussian process prior is able to fit this “threshold-induced” preference on the outside good.

In a fourth study, we use an outside good functional form $u(z) = 0.7z$, which displays no satiation (Figure 2.6). Once more, the parametric functional form struggles to fit the change in satiation rate and compromises across the two rates. On the other hand, the Gaussian process prior is able to fit this “threshold-induced” preference on the outside good.

Table 2.7, Table 2.10 and Table 2.13 show how each model (parametric baseline and Gaussian process prior model) generalize on test observations. We use the posterior predictive modal values to generate predictions for each final good (0,1,2). The hit probabilities are higher for each Gaussian process specification, and the negative log predictive densities are also lower, suggesting that the GP-based models fit the data better.



(a) Parametric

(b) Gaussian process

FIGURE 2.4: Estimated utility of the outside good when the misspecified parametric functional form $u_z(z) = \alpha * \log(z)$ is imposed (a) and when a Gaussian process prior is used on the latent utility function (b). The ground truth is a utility function of the following form: $u_z(z) = \sin(0.7z) + 0.7z$. Note that the realized utility values are unobserved by the analyst. The hyperparameters of the Gaussian process are manually set and not estimated, for simplicity and as it significantly reduces the computational burden.

f

	Predicted		
Actual	0	1	2
0	0.71	0.16	0.08
1	0.05	0.86	0.18
2	0.00	0.07	0.88

TABLE 2.5:
Nonhomothetic
Hit Probability: 0.82
NLPD: 86.06

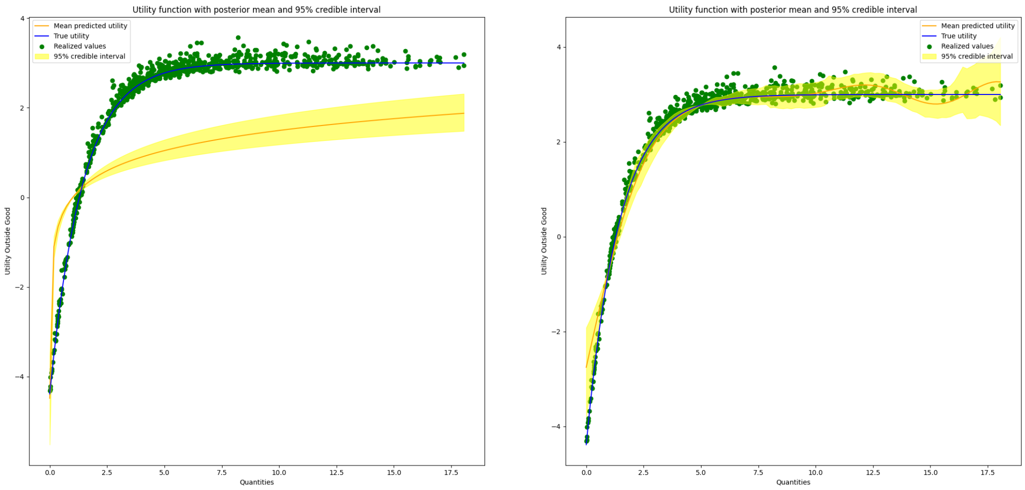
	Predicted		
Actual	0	1	2
0	0.71	0.16	0.08
1	0.00	0.86	0.24
2	0.00	0.02	0.96

TABLE 2.6:
Gaussian process
Hit Probability: **0.84**
NLPD: **81.61**

TABLE 2.7: Confusion matrix on test observations, comparing a nonhomothetic model and a Gaussian process model, with a ground truth of $u_z(z) = 3 - \exp(2 - 0.7z)$ and a misspecified parametric prior of $u_z(z) = \alpha \log(z)$. Hit probabilities are calculated using posterior predictive modal values. The Gaussian process model demonstrates improved predictive accuracy and a better fit to the test data in terms of negative log predictive density (NLPD).

2.5 Discussion

To be able to use the proposed model in both applications, a regime of informative likelihood is necessary, such that the functional form is identified by the data. In particular, a high density of the data through multiple price points changes is crucial to identify the functional form of utility. Another limitation in the first application is the joint Gaussian process prior on marginal utility and its derivative function does not



(a) Parametric

(b) Gaussian process

FIGURE 2.5: Estimated utility of the outside good when the misspecified parametric functional form $u_z(z) = \alpha \log(z)$ is imposed (a) and when a Gaussian process prior is used on the latent utility function (b). The ground truth is a utility function of the following form: $u_z(z) = 3 - \exp(2 - 0.7z)$. Note that the realized utility values are unobserved by the analyst. The hyperparameters of the Gaussian process are manually set and not estimated, for simplicity and as it significantly reduces the computational burden.

	Predicted		
Actual	0	1	2
0	0.77	0.09	0.12
1	0.15	0.86	0.18
2	0.02	0.06	0.79

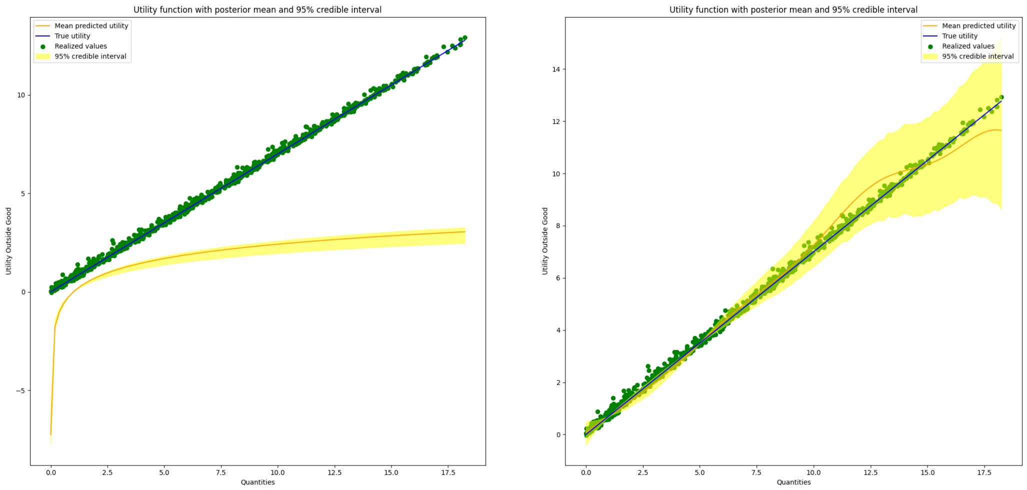
TABLE 2.8:
Nonhomothetic
Hit Probability: 0.82
NLPD: 91.24

	Predicted		
Actual	0	1	2
0	0.85	0.08	0.00
1	0.03	0.95	0.09
2	0.03	0.05	0.79

TABLE 2.9:
Gaussian process
Hit Probability: **0.895**
NLPD: **54.52**

TABLE 2.10: Confusion matrix on test observations, comparing a nonhomothetic model and a Gaussian process model, with a ground truth of $u_z(z) = \sin(0.7z) + 0.7z$ and a parametric prior of $u_z(z) = z^\alpha$. Hit probabilities are calculated using posterior predictive modal values. The Gaussian process model demonstrates a significantly higher predictive accuracy and a better fit to the test data as indicated by the lower NLPD value.

have any monotonicity restrictions. Economic theory suggests that the utility function must be quasiconcave to have necessary and sufficient first-order Kuhn-Tucker conditions. When utility is additive, and we have $U(q_1, \ldots, q_n) = F(\sum_{k=1}^{K} u_k(q_k))$ where F is a strictly increasing real function, then at least $n-1$ of the functions $u_1, \ldots, u_n$ must be concave functions, without necessarily destroying the quasiconcavity of U (Yaari, 1977). Van Soest, A. Kapteyn, and Kooreman (1993) show that



(a) Parametric

(b) Gaussian process

FIGURE 2.6: Estimated utility of the outside good when the misspecified parametric functional form $u_z(z) = \alpha \log(z)$ is imposed (a) and when a Gaussian process prior is used on the latent utility function (b). The ground truth is a utility function of the following form: $u_z(z) = 0.7z$. Note that the realized utility values are unobserved by the analyst. The hyperparameters of the Gaussian process are manually set and not estimated, for simplicity and as it significantly reduces the computational burden.

	Predicted		
Actual	0	1	2
0	0.83	0.10	0.07
1	0.06	0.91	0.19
2	0.01	0.09	0.63

TABLE 2.11:
Parametric baseline
Hit Probability: 0.845
NLPD: 69.83

	Predicted		
Actual	0	1	2
0	0.83	0.11	0.04
1	0.06	0.92	0.15
2	0.03	0.05	0.74

TABLE 2.12:
Gaussian process
Hit Probability: **0.865**
NLPD: **59.06**

TABLE 2.13: Confusion matrix on 100 test observations, with a ground truth of $u_z(z) = 0.7z$ and a misspecified parametric prior of $u_z(z) = \alpha \log(z)$. NLPD: negative log predictive density. Hit probabilities are calculated using posterior predictive modal values. The Gaussian process model predicts better on test data, misclassifies choice instances less often, and fits the test data better.

when a maximum likelihood estimator is used and utility is not strictly quasiconcave, the Kuhn-Tucker conditions may suffer from the "coherency problem" that is they may not yield a unique vector of optimal quantities. However, they also mention that flexible and tractable demand systems only have local concavity properties, and coherency can be guaranteed by imposing regularity conditions in some relevant region of price or quantity space. One way to make impose a monotonic prior on marginal utility is to use a transformed Gaussian process. We refer the interested readers to Appendix B.2 for an original method to implement this monotonic prior.

The second application relaxes the assumption of strong separability by using a nonhomothetic choice model with intermediary and final goods, building on household production theory. The final goods subutility is separable, but the intermediary goods utility becomes non-separable. However, the second application makes the assumption of a known input-output matrix, where the exact mix of intermediary goods is known ahead of time.

2.6 Conclusion

This chapter proposes a general framework to incorporate a flexible functional prior on direct utility models, with a structural interpretation. We highlight that misspecifying the functional form of the outside good severely impacts the estimation of preference parameters for inside goods demanded in continuous quantities, and in the case of final goods demanded in discrete quantities. The first application displays rich insights on baseline preferences and satiation rates that a parametric model could not capture adequately. The flexibility of Gaussian processes is a desirable property to obtain additional information on preferences at the individual level. The second application develops a nonhomothetic choice model that overcomes the limiting assumption of additive utility, through household production theory. We show in a series of numerical simulations, that our Gaussian process choice modeling framework is robust to misspecification in the outside good utility functional form, which in turns leads to better choice prediction, and more accurate price effect estimation.

Chapter 3

Digital Twins: A Generative Approach for Counterfactual Customer Analytics

This chapter provides a novel methodology, Digital Marketing Twins, that automatically extracts latent features from individual-level brand survey responses to inform a statistically-principled, deep generative model of customer-side brand affinity and firm-side performance factors. The proposed model enables marketers to find drivers of individual-level brand affinity, as opposed to traditionally observed metrics that must be analyzed in aggregation. The framework serves a counterfactual purpose at the customer level. The generative part of the model *completes* the distribution of survey responses over time, and across firms – thereby addressing the archetypal missing data problem – by imputing customer responses in counterfactual regimes. The proposed prescriptive framework also proposes policy optimization through customer surveys, using Bayesian optimization, which efficiently identifies "paths of least resistance" among customer responses to service-quality questions – a search that otherwise would represent a complexity of $\mathcal{O}(n^d)$. This research applies Digital Marketing Twins methodology to the competitive landscape of the U.S. wireless telecommunications retail market, leveraging a unique dataset of large-scale quarterly brand surveys from all three major carriers (AT&T, T-Mobile, and Verizon) from 2020 to 2022. Empirically, this approach reveals latent asymmetries in competition in terms of brand affinity, together with a nonlinear increase in brand affinity for certain types of drivers, such as satisfaction with network speed, but a nonlinear decrease in brand affinity for customers who report greater likelihoods of changing plans, providers, or devices, relative to their current wireless services.

3.1 Introduction

Customer surveys are ubiquitous tools. Marketers leverage them to fuel brands and boost corporate growth, as well as determine the causes of customer satisfaction and of customer churn. Marketing researchers adopt them to learn customer preferences, gauge customer satisfaction, identify competitive offers, improve existing products and services, tailor marketing strategies, and innovate personalized services. Due to the ever growing complexity, and frequency of customer surveys and survey touchpoints though, firms today mostly rely on third-party platforms (e.g., Salesforce, HubSpot) or survey companies (e.g., Ipsos and Kantar) to execute large-scale customer surveys, known commercially as *brand surveys*. The increasing scale and scope of such brand surveys also is part of a broader trend of marketers embracing data-driven approaches, with an emphasis on the use of behavioral metrics to compute customer lifetime value (CLV) (e.g., Venkatesan and Kumar, 2004; Fader, Hardie, and K. L. Lee, 2005). Behavioral metrics help identify which customers are at risk of churning, though recent research also calls for efforts to distinguish the causes from the predictors of churn (Braun and Schweidel, 2011; Ascarza et al., 2018; Ascarza, 2018). That is, predictors of churn include demographics and behavioral patterns that statistically indicate a high likelihood of discontinuing products or services, but the causes of churn might include poor customer service, high prices, product quality issues, or more attractive offerings from competitors. Customer satisfaction targets the root issues that lead to customer attrition (Gustafsson, Johnson, and Roos, 2005). Moreover, customer satisfaction and dissatisfaction drive companies' stock prices asymmetrically: dissatisfaction harms returns far more than a one-unit increase enhances them (Malshe, Colicev, and Mittal, 2020). Accordingly, a new methodology is needed to carve out the "paths of least resistance" to individual-level customer satisfaction; this research proposes a prescriptive framework that pioneers such policy optimization by relying on customer surveys.

Despite the numerous benefits of customer surveys for marketers, there are two issues – one theoretical and the other practical – that plague most large-scale surveys carried out for customer relationship management (CRM) purposes. First, the surveys are difficult to integrate into prescriptive frameworks. Linear functional forms, strong parametric assumptions, and limited consideration of customer heterogeneity as a result of limited or incomplete individual-level data across brands makes it difficult for marketers to understand customer churn and retention in mature, competitive environments. Which marketing action should be recommended when customer satisfaction declines? Although one obvious lever is promotional offers, there are others; for example, in the wireless telecommunications industry, managers can increase customer satisfaction by improving network quality and network speeds, providing better data plans, strengthening brand perception, and solving problems that customers encounter when using providers' devices and services. However, the question remains: which aspects should be prioritized at the customer level? Second, in practice, customer surveys often are repeated cross-sectional. Because repeated cross-sectional surveys represent *different* sets of customers at various points in time, they cannot track individual changes. Unlike longitudinal surveys, they take “snapshots” and thereby provide less depth of information about individual respondents. If the sampled population changes significantly over time, comparisons between different cross-sectional surveys become challenging, if not impossible. For this reason, the nature of data variations in the repeated cross-sectional format of brand surveys is described as *pseudo-longitudinal*.

In this chapter, I propose a novel methodology – Digital Marketing Twins – that leverages large-scale brand surveys conducted by a focal firm and its competitors in the U.S. wireless telecommunications retail market. This methodology finds paths of least resistance to individual-level customer satisfaction, in a statistically principled way. It uses a unique dataset from a representative sample of customers of AT&T, T-Mobile, and Verizon, the three major players in the U.S. telecommunications market. Using quarterly cross-sectional survey responses that span ten quarters – from 2020 to 2022 – the methodology overcomes both the substantive and technical limitations previously mentioned. The framework builds a generative model of customer preferences by flexibly mapping individual-level surveyed characteristics to various dimensions of customer satisfaction. Generative models capture the joint probability distribution between observed and latent variables of interest. In practice, they provide the steps that explain how the data are assumed to be generated, allowing marketers and researchers to incorporate domain expertise into their models. The generative aspect not only supports the forecasting of customers’ responses in the next quarter, but also provides counterfactual responses according to different scenarios, such as customer responses as if they were using a different wireless carrier, all else being equal.

The digital twins approach, already an established method for counterfactual simulations in the realm of manufacturing, presents a novel and previously unexplored avenue for application within the marketing field. Until now, this innovative approach has, to the best of my knowledge, remained unapplied to this context. Digital twins integrate data from different sources to mimic the behavior of physical objects or systems; they can be used to test hypotheses, simulate scenarios, and optimize the performance of the systems. The use of generative models as a basis for digital twins is not novel; for example, generative adversarial networks (GANs) and conditional GANs can learn distributions of interest in structures with material nonlinearities and uncertainties (Tsialiamanis et al., 2021). In marketing contexts, digital twins serve a counterfactual purpose at the customer level. The generative part of the model *completes* the distribution of survey responses over time, and across firms, such that it can address the archetypal missing data problem. The proposed Digital Marketing Twins methodology offers a solution to the missing data problem that arises from the pseudo-longitudinality of brand surveys – in which individual respondents can be only observed in one time period and at one company – by imputing customer responses in the next time period and in a counterfactual regime in which all individually observed characteristics remain constant. The goal is to infer customer satisfaction under counterfactual regimes of their experiential, engagement, and usage characteristics, identifying the potential causes of satisfaction.

Table 3.1 provides a summary of the conceptual benefits of the Digital Marketing Twins framework. Quasi-experimental methods can be particularly useful for understanding the effect and strategic value of an intervention on an outcome of interest from recorded CRM data. A combination of propensity score matching and a flexible Bayesian parametric or nonparametric model such as a GAM, a Gaussian process (GP), or a mixture-of-normals can be useful for making counterfactual predictions. However, these methods are seldom scalable out-of-the box; they require careful selection of kernel functions and/or hyperparameters, which in turn requires expert knowledge. The quasi-experimental methods' inference (especially of natural experiments) is necessarily to quantifying what *has* happened, rather than what *could* happen. Although state-of-the-art predictive CRM methods provide forward-looking analytical insights, often with the help of rich machine learning techniques, they lack the ability to provide counterfactuals by focusing on predictors of customer churn, neglecting complex structures such as competitive effects, and relying on flexible statistics of customer behavior.

	Predictions / Tactical	**Counterfactuals / Strategic**
Retroactive	Basic CRM Reports	Quasi-Experimental Methods
Proactive	Predictive CRM Models	Digital Twins

TABLE 3.1: Digital Marketing Twins as a proactive and strategic framework for customer analytics.

From a technical standpoint, the goal of this research is to develop a novel deep generative and probabilistic latent factor model, as well as to leverage Bayesian optimization to find the best marketing actions to recommend at the individual level, from a large-scale survey. The Digital Marketing Twins model captures customer-side brand affinity at the individual level, for each brand, in each time period, controlling for observed heterogeneity and firm-side factors. The inference model mapping from data to the latent space is parameterized by a neural network, for high flexibility. Latent, customer-side brand affinity provides an interpretable layer that maps to a latent utility model that in turn yields an ordinal logit structure for brand survey questions pertaining to customer satisfaction. To generate counterfactual responses and missing quarters, it relies on amortized inference, learning a set of parameters that can map any data point to the latent space. I train the model using stochastic variational inference with mini-batching, for high scalability and uncertainty quantification. After training, a Bayesian optimization method maximizes individual-level, latent, customer-side brand affinity for customers of the focal firm, to discover the marketing actions most likely to increase customer satisfaction. This maximization leads to a path of least resistance at the individual level, enabling marketers to use surveys to identify causes of satisfaction. Applications of Bayesian optimization on a latent space has been applied to other contexts, outside marketing (e.g., Gómez-Bombarelli et al. 2018; Griffiths and Hernández-Lobato 2020), and is useful in situations in which gradients are not accessible.

The remainder of the chapter is organized as follows: Section 2 presents a literature review. In Section 3, contains a description and exploration of the data, using simple descriptive techniques. Section 4 provides the methodology. Section 5 provides fit metrics, benchmarks the Digital Marketing Twins model against nested baseline models, analyzes the probabilistic latent factors, and provides counterfactual results. Section 6 introduces a novel prescriptive framework to optimize customer satisfaction with Bayesian optimization. Section 7 concludes.

3.2 Related Literature

This work contributes to academic literature on digital twins using generative modeling, as well as to literature on machine learning methods in marketing for competitive environments; it also shows that early models for customer satisfaction are special cases of the proposed framework.

3.2.1 Digital Twins and Generative Modeling

Tsialiamanis et al. (2021) suggest how to advance system simulation by creating digital twins for specific systems, referring to fields such as manufacturing, control systems, the Internet of Things (IoT), smart cities, social networks, and management. Digital twins can help predict the behavior of structures in different situations, thus maximizing the operational lives of the structures and minimizing costs. However, the construction of digital twins is inherently complex and uncertain. Aleatory uncertainty, related to random events, and epistemic uncertainty, related to a lack of knowledge, are key considerations. To address these issues, Tsialiamanis et al. (2021) propose the use of generative models as the foundation for a digital twin, providing estimations of aleatory and epistemic uncertainty. They study two types of generative models: the Stochastic Finite Element (SFE) method – a physics-based, white-box model – and the Conditional Generative Adversarial Network (cGAN) – a data-driven, black-box model. Each has strengths and limitations. For example, SFE models excel in predefined conditions but struggle with unknown scenarios, whereas cGANs can perform across a wide range of conditions but cannot extrapolate beyond available data. With a hybrid, grey-box approach, incorporating both models to overcome these limitations, generative models might better accommodate uncertainty in digital twins. By combining a generative white box (SFE) and a generative black box (cGAN), they propose a fully generative grey box that they assess in relation to other existing models, such as variational auto-encoder (VAE) and

Gaussian processes (GP). In this chapter, on the other hand, I use a deep generative model building on a variational auto-encoding neural network to mirror the competitive environment, and use a Gaussian process to optimize the latent customer-side brand affinity.

M. G. Kapteyn, Pretorius, and Willcox (2021) propose a new mathematical foundation for digital twins, that is, computational models that mirror structures, behaviors, and contexts of physical assets. Because digital twin applications usually require extensive resources and expertise for implementation, the authors propose a unifying mathematical model that uses dynamical systems theory and probabilistic graphical models, with the digital twin and the physical asset modeled as coupled dynamical systems that evolve over time, and the digital twin constantly updating its internal models according to observational data. By dmonstrating this approach with a digital twin of an unmanned aerial vehicle (UAV), they show how the model aids in calibration by updating internal models, and facilitating decision-making. They conclude with the presentation of an abstract state-space formulation for digital twins, describing a realized dynamic decision network based on this mathematical model and illustrating its application to a UAV's structural digital twin.

Finally, Yu et al. (2021) propose a health monitoring solution for complex systems in smart manufacturing, applying a digital twin approach with a nonparametric Bayesian network model. With advancements in sensor technology and artificial intelligence, modern manufacturing systems need to be intelligent, visual, and capable of self-assessing their health throughout their life cycle. The Prognostic and Health Management (PHM) process is crucial in this context, and the proposed model offers an innovative solution for tracking the health states of such systems. The model collects sensor data from the physical world, updating its simulated physical model in real time and providing optimization and decision support. Their nonparametric Bayesian network model can adapt in real-time too, thus reducing model uncertainty. Yu et al. (2021) also include model validation experiments on electro-optical systems, and provide more accurate health monitoring than a traditional data-driven Convolution Neural Network (CNN) approach.

3.2.2 Machine Learning Methods in Marketing for Competitive Environments

Among the proposals for machine learning techniques to study market structures and competitive landscapes, Netzer et al., 2012 systematically analyze online user-generated content to "listen" to what customers write about a focal firm's and competitors' products; they use text mining to overcome the difficulties involved in extracting and quantifying the wealth of online data that customers generate and network analysis tools to convert the mined relationships into co-occurrences among brands or between brands and terms. (Tirunillai and Tellis, 2014) extract latent dimensions of customer satisfaction with quality, using an unsupervised latent Dirichlet allocation model, and T. Y. Lee and Bradlow, 2011 automatically elicit product attributes and extract brands' relative positions from online customer reviews, providing both predictive and descriptive support for managerial decision making.

Brand competition occurs not only in single markets, but also in different submarkets and structured markets. The hypothesis of multiple structured markets (Kannan and Wright, 1991) helps us understand how brands compete by including marketing mix variables. In type-primary markets, "switchers" are highly responsive to changes in marketing mix variables whereas in brand-primary markets, the "loyal" segment remains relatively unresponsive to marketing programs (e.g., in contexts of ground coffee purchases or store panel records). Ringel (2023) recently proposed the visualization of brand competition in a multimarket membership product (MMP) context, in which products that compete in multiple submarkets that are each characterized by distinct competitors and customer preferences, with competitive relationships inferred from customers' online searches using bootstrapped neural network product embeddings in the digital camera market.

3.2.3 Customer Satisfaction and Survey Research

Using an ordered probit model, Kekre, Krishnan, and Srinivasan (1995) study determinants of customer satisfaction for software products and service support for mainframes and workstations. Their main dependent variable is an overall satisfaction score, measured on an ordered categorical scale. The authors can explain how certain features of the software, such as reliability, capability and usability, affected overall satisfaction. They consider other explanatory variables, such as the type of product and the type of user, allowing for interaction effects. If an ordered logit were substituted for their ordered probit, their model can be nested within the proposed framework, by replacing amortized neural networks with linear functions, assuming that business key performance indicators (KPI) have no impact on customer satisfaction and considering only overall satisfaction as a unique target variable.

3.3 Exploring the Data

The data for this study consist of repeated cross-sectional responses from a brand survey for all major U.S. telecom carriers (AT&T, T-Mobile, and Verizon) between the third quarter of 2019 and the third quarter of 2022. Responses are recorded quarterly. For each carrier – not necessarily in the order previously mentioned, for confidentiality – I observe responses from a sample of 8770 customers, 7129 customers, and 4370 customers, in every quarter. The number of customers is the same across quarters for a given carrier, but the sample of customers for each carrier differs between quarters (i.e., repeated cross-sectional data).

According to managerial sources from one of the three major U.S. telecom carriers, the objectives of this survey were threefold: (1) gain an understanding of how wireless, internet, and pay TV customers view and rate customer experience with their carrier or provider, (2) determine the driving factors of customer satisfaction, and (3) determine what the focal firm does well and where it falls behind competitors, according to not only customer satisfaction (i.e., net promoter score) but also specific drivers and attributes.

3.3.1 Inputs and Outputs

The questions in the survey data fall into three categories, representing three different goals. The first group of questions provides customer characteristics; their characteristics have a predictive function, because they cannot be manipulated by managers (e.g., age and ethnicity cannot be influenced by any marketing action). The second group of perceptual questions offer immediate strategic value to managers, in that they ask customers about their feeling toward competing carriers. The third group of questions relates directly to customer satisfaction and form the basis for the proposed digital twin approach, because I assume that managers aim to maximize customer satisfaction. Therefore, the survey questions reflect three categories:

- Predictive Variables: during the inference phase, and at test time, these variables are fixed. In the optimization phase, they remain fixed. A key assumption of the model is that invariant predictors completely characterize customer heterogeneity. For the empirical application, I use large numbers of socio-demographic and usage questions, including age, gender, race or ethnicity, annual household income before taxes, devices at home, name of the wireless service provider, type of plan, tenure with provider, dollar amount paid per month for the plan, data usage, 5G usage, and rewards program.

- Strategic Variables: at training time and test time, these variables are fixed. At optimization time, these variables are the arguments of the optimization problem, and are assumed to be manipulated by the marketing analyst. They include:

  - Importance of any of the following according to likelihood to recommend: network in rating, price / value; billing; customer service; general feeling; plans; rewards and benefits; other factors.
  - Satisfaction with network speed; network reliability; data plans that meet my needs; value of price paid; accuracy of billing; rewards and recognition; ease of doing business; solving problems for the first time; "brand for me"; total cost of wireless service; device selection.

- Target Variables: these variables are reconstructed at inference time, and predicted at test time, and optimized at optimization time. They include[1]:
  - Likelihood to recommend (LTR) (0-10);
  - Likelihood to recommend current provider's phone to a friend or a colleague (Phone LTR) (0-10);
  - Likelihood to switch wireless service providers within the next 12 months (Intention to Switch) (0-4);
  - Overall satisfaction with current provider (0-9);
  - Overall feeling about current provider (0-4);
  - Overall feeling about competitions' providers[2] (0-4).

The model also controls for different aspects of firm performances, using Generally Accepted Accounting Principles (GAAP) and non-GAAP measures published quarterly by AT&T, T-Mobile, and Verizon between the second quarter of 2020 and the fourth quarter of 2022. For brevity, I denote these variables as business key performance indicators (KPIs). They include total revenue, operating revenue, cost of revenue, gross profit, operating expense, churn, and average revenue per user (ARPU). The measures are standardized. Tables C.1 and C.2 (in the Appendix) lists all questions included as target variables and strategic variables. Figure C.1 includes summary statistics at the question level, per carrier.

3.3.2 Multivariate Analysis

Before providing a generative model of the target variables (LTR, Phone LTR, Satisfaction, Overall Feeling about Carrier {1, 2, 3}, Intention to Switch), it is helpful to understand how they are associated with one another. Therefore, I undertake a correlation analysis of the target variables, at the carrier level.

First, I recode the Intention to Switch as Retention Likelihood, applying the formula $f(x) = 4 - x$. Intuitively, Satisfaction, Likelihood to Recommend and Overall Feeling about Own Provider should correlate positively with Retention Likelihood; an empirical analysis verifies these correlations (Figure 3.1). For each carrier, the correlation between retention likelihood and satisfaction ranges from 0.28 to 0.31. The correlation between Retention Likelihood and Feeling about Current Provider also is positive (respectively, 0.50, 0.43, and 0.46 for Carriers 1, 2, and 3). Phone LTR is also positively correlated with Retention Likelihood. The strong positive correlations between LTR and Feeling about Own Provider and Satisfaction suggest that

[1] The complete list of strategic and target variables is available in Appendix A; because there are more than 300 one-hot encoded predictors, they are not listed here. The complete list remains available upon request.

[2] For example, an AT&T customer is asked about their overall feeling about T-Mobile and Verizon, as two separate questions.

marketers at least indirectly capture a measure of satisfaction when they record the popular Net Promoter Scores[3] (F. F. Reichheld, 2003).

It is more challenging to understand the relationship between Feeling about Competitions' Providers and other target variables. More positive feelings are associated with a lower Retention Likelihood (correlations from -0.08 to -0.18, Figure 3.1). However, more positive feelings are also associated with higher LTR, Phone LTR, and Satisfaction, suggesting that customers may simply be "happier" about the telecommunications industry in general. This suggestion is corroborated by the slightly positive correlation between all Feeling measures about Own and Competitor's providers.

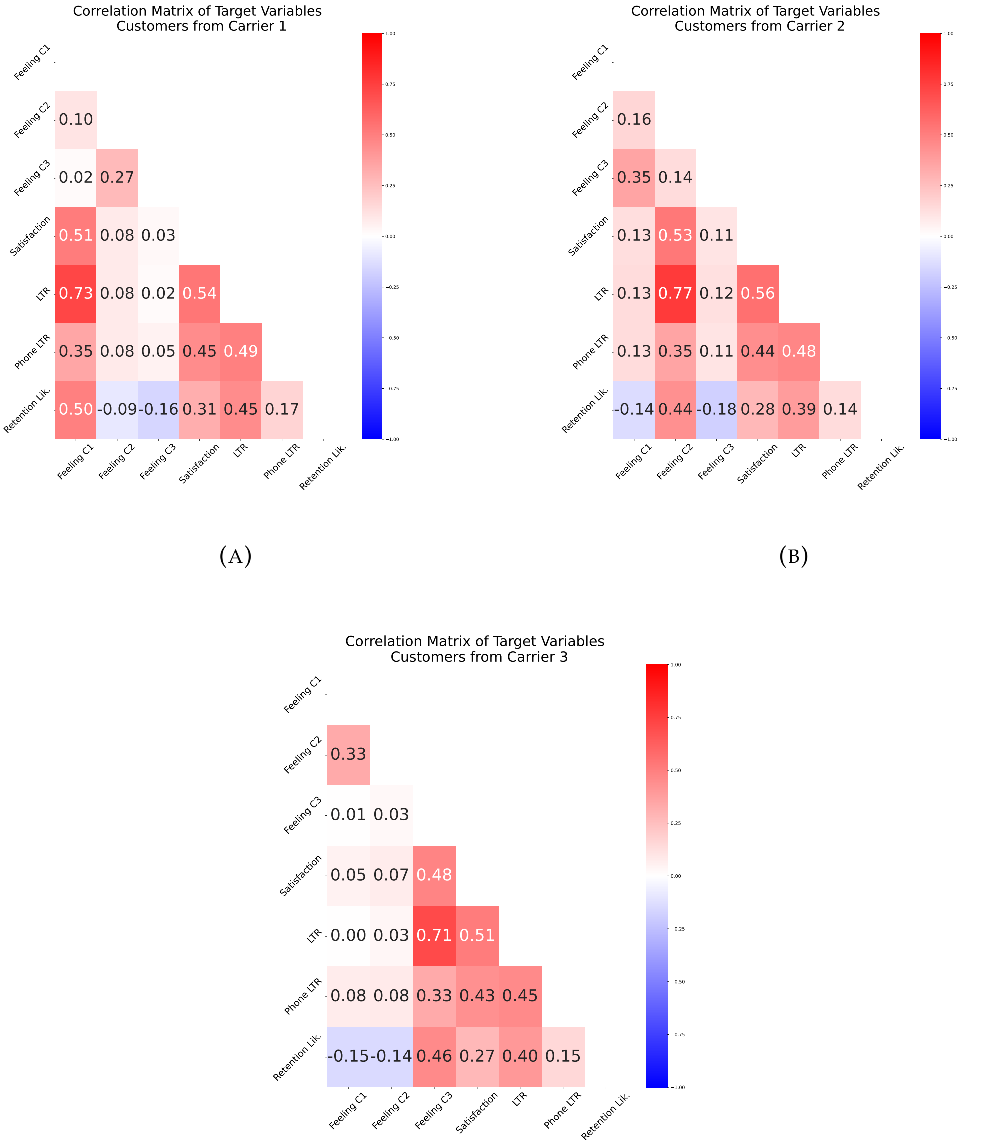
Finally, it is interesting to notice symmetries and asymmetries between the three carriers in terms of correlations across target variables. In terms of symmetries, customers from all carriers tend to have more positive or more negative feeling about all competitions' carriers simultaneously; moreover, the variable that correlates most with Retention Likelihood is Feeling about Own carrier, whereas Overall Satisfaction comes second. In terms of asymmetries, customers from Carrier 3 do not express positive or negative associations with their Feeling about Own Carriers and Competitions' Carriers (correlations of 0.01 and 0.03, in Figure 3.1) whereas customers from Carrier 1 and especially Carrier 2 tend to have a stronger associations (correlations of 0.14 and 0.16 for Carrier 2, Figure 3.1).

3.3.3 Linear Modeling is Limited for Analyzing Brand Surveys

The first step in gauging whether a nonlinear model is needed to investigate the relationship between explanatory variables (invariant predictors) is to compare two simple discriminative machine learning models. For simplicity, I use multiple output linear regression as a benchmark for the discriminative linear model, and multiple output random forest as a benchmark for the discriminative nonlinear model.

Multiple output linear regression is a generalization of the simple linear regression model when more than one output variables is considered. The model learns a linear relationship between the input variables and each of the output variables. Each output variable is modeled as a linear combination of the input variables plus an error term. In contrast, random forests are ensemble learning methods that operate by constructing a multitude of decision trees and outputting the mean prediction for a regression task. A multiple output random forest is an extension of this technique to handle multiple output variables. This method is useful when the output variables are not independent of each other and share some correlation, as indicated in Subsection 3.2.

[3] The Net Promoter Score is the measure that transforms LTR by assigning a +1 to respondents who indicate a LTR of 0 to 6, 0 to those who indicate a LTR between 7 and 8, and +1 to those who provide a LTR of 9 and 10, then taking the average across all respondents.


Correlation Matrix of Target Variables
Customers from Carrier 1

(A)

Correlation Matrix of Target Variables
Customers from Carrier 2

(B)

Correlation Matrix of Target Variables
Customers from Carrier 3

(C)

FIGURE 3.1: Correlation matrices for target variables in the survey, for each carrier.
Notes: Feeling C1, Feeling C2, and Feeling C3 respectively refer to Overall Feeling about Carrier 1 (a), Carrier 2 (b), and Carrier 3 (c), respectively.

Fitting the multiple-output linear regression and the multiple-output random forest on the pooled data shows that the coefficient of determination score R^2 is higher for all target variables in the random forest analysis (Table 3.2). This higher goodness-of-fit metric indicates that the nonlinear model better captures the relationships between input (invariant and strategic) variables and target variables.

Although linear regression and random forest are useful models, they are fundamentally predictive and do not necessarily provide meaningful interpretations of the relationships between inputs and outputs. For counterfactual reasoning, other techniques may be more desirable, such as structural equation modeling (SEM). Yet, even if SEM may be more interpretable, it also yields questionable assumptions, because it is difficult to know *a priori* how the different input variables relate to latent constructs of interest that explain the various target variables that managers care about and try to optimize.

Data Source	**Multiple Output Discriminative Models**	
	Linear Regression (Test R^2)	**Random Forest (Test R^2)**
Q12 (Likelihood-to-recommend, LTR)	0.78	**0.82**
Q27 (Phone LTR)	0.30	**0.42**
Q18 (Satisfaction)	0.58	**0.67**
Q20Ar1 (Overall Feeling Carrier 1)	0.38	**0.58**
Q20Ar2 (Overall Feeling Carrier 2)	0.40	**0.57**
Q20Ar3 (Overall Feeling Carrier 3)	0.39	**0.55**
Q20 (Intention to Switch)	0.52	**0.60**

TABLE 3.2: Goodness-of-fit of two discriminative models on a test sample of the data.
Notes: Multi-output Linear regression is a standard benchmark, while multi-output random forest is a nonlinear benchmark. A nonlinear model better explains the variation in the data.

3.4 Modeling Framework

Subsection 3.4.1 details the model architecture and training, based on mapping customer - side and firm - side input variables to latent variables using amortized neural networks. After documenting the latent variables and their marketing interpretations (Subsection 3.4.2), Subsection 3.4.3 presents the model layer for ordered categorical variables. Subsection 3.4.4 details the inference procedure, with implementation details. Section 3.5 outlines the predictions tasks, and Section 3.6 offers a description of the optimization phase.

3.4.1 Mapping Customer-Side and Firm-Side Input Variables to Latent Parameters through Amortized Neural Networks

The survey data refer to N respondents and K carriers. In a single quarter t, for a given firm k, a subset N_{kt} of these N respondents are surveyed, such that $\sum_{k=1}^{K} \sum_{t=1}^{T} N_{kt} = N$. The number of respondents within a firm remains constant over time, but individuals are not surveyed more than once, such that $N_{k1} = \cdots = N_{kT}$

for all $k = 1, \dots, K$. Throughout this chapter, $K = 3$, referring to the three largest carriers: AT&T, T-Mobile, and Verizon[4]. The survey data has J questions. I label the J^{pred} predictive variables $\mathbf{x}_{it}^{pred}$, the J^{str} strategic variables $\mathbf{x}_{it}^{str}$ and the $J^{targets}$ targets $\mathbf{y}_{ijtk}$, where $i = 1, \dots, N$ customers, $j = 1, \dots, J$ questions, $k = 1, 2, 3$ firms. In summary, $J^{str} + J^{pred} + J^{targets} = J$. The training phase does not make a conceptual distinction between invariant predictors and strategic variables, because they enter the same neural network. Therefore, $\mathbf{x}_{it} = [\mathbf{x}_{it}^{str}, \mathbf{x}_{it}^{pred}]^T$. For a given quarter t and a given firm k, business KPI are $\mathbf{x}_{kt}^{KPI}$, which is a vector of size H.

Amortized inference refers to inference over variational parameters that are parameterized by a function of the data, instead of approximating separate variables for each data point (Cheng Zhang et al., 2018). For this research, the parameterized function is the neural network $f(.)$ that represents the parameters of the variational distribution across all data points from the J questions in the survey. An alternative would be to separately learn a set of parameters for each data point, rather than learning a set of mean and location parameters for each customer, at each time period, and each firm. The word "amortized" herein means that the cost of learning the variational parameters is "amortized" over all the data points.

Amortized inference is a powerful way to infer the posterior over customer-level and firm-level latent variables according to $\mathbf{x}^{KPI}$, $\mathbf{x}^{str}$ and $\mathbf{x}^{pred}$. Using variational inference to approximate the posterior distribution of customer-side and firm-side latent variables implies replacing the locational variational parameters with a function of the data where parameters – weights and biases of neural networks – are shared across all data points, for all firms and at all quarters. The neural network parameters automatically learn a complex representation of the inputs across firms and over time, and this representation is mapped to latent variables that are the building blocks of the target variables. As a major methodological contribution, the current study proposes amortized inference as a way to augment repeated cross-sectional data.

The feed-forward neural networks $f(.)$ and $g(.)$ map customer-side and firm-side, respectively, input variables to a set of latent location and scale parameters that generatively model the target variables. In such feed-forward neural networks, hidden layers are dense and sequentially connected. Consider the feed-forward neural network function $f(\mathbf{x}; \theta_f)$ with D hidden layers; is detailed as follows. The input layer is $d = 0$, hidden layers are $d = 1, 2...$, and the output layer is D. The weights connecting layer d and layer $d + 1$ can be referred to as $W^{(d)}$, and the biases in layer $d + 1$ are indicated by $b^{(d)}$. The pre-activation at layer $d + 1$ can be denoted as $a^{(d+1)}$, and the post-activation is $h^{(d+1)}$. The activation function is the hyperbolic tangent

[4] Sprint was also a major carrier but merged with T-Mobile U.S. on April 1, 2020. It was the fourth-largest telecommunications carrier in the United States before the merger. Since the data starts in Q2 2020, I do not consider Sprint customers in the analysis.

($tanh$), such that:

$$\mathbf{a}^{(1)} = W^{(0)}\mathbf{x}_{it} + \mathbf{b}^{(0)} \tag{3.1}$$

$$\mathbf{h}^{(1)} = \tanh(\mathbf{a}^{(1)}) \tag{3.2}$$

$$\mathbf{a}^{(2)} = W^{(1)}\mathbf{h}^{(1)} + \mathbf{b}^{(1)} \tag{3.3}$$

$$\vdots$$

$$\mathbf{a}^{(D)} = W^{(D-1)}\mathbf{h}^{(D-1)} + \mathbf{b}^{(D-1)} \tag{3.4}$$

$$\mathbf{h}^{(D)} = \tanh(\mathbf{a}^{(D)}) \tag{3.5}$$

$$\mu_{ikt} = W_{\mu}^{(D)}\mathbf{h}^{(D)} + \mathbf{b}_{\mu}^{(D)} \tag{3.6}$$

$$\nu_{ikt} = \exp\left(W_{\nu}^{(D)}\mathbf{h}^{(D)} + \mathbf{b}_{\nu}^{(D)}\right) \tag{3.7}$$

Here, $\mathbf{x}_{it}$ is the batch input to the network. Because ν_{ikt} is a variance and must be non-negative, I apply an exponential function to obtain it from $\mathbf{a}^{(D)}$. The weights and biases (collectively referred to as $\boldsymbol{\theta}_f$) are learned by training the network. These weights and biases are parameters of amortized neural networks.

For the feed-forward neural network $g(\mathbf{x}^{KPI}; \boldsymbol{\theta}_g)$ with D' hidden layers, the inputs are the KPI for the three major carriers in the U.S. market (AT&T, T-Mobile, and Verizon), published quarterly over the corresponding 10 quarters of survey data. The neural network's output at a batch level is a concentration parameter γ_{ktl} and a rate parameter ω_{ktl}. The function g also relies on amortization to learn a shared representation across all quarters and firms instead of learning individual γ_{ktl} and ω_{ktl}. The dimension l refers to a set of L latent dimensions summarizing the various aspects of performance across firms and over time. These L latent dimensions provide dimensionality reduction, similar to principal components in principal component analysis.

Finally, the use of the *tanh* activation function introduces non-linearities between layers, allowing the network to learn complex mappings from inputs to outputs. The *tanh* function is particularly well-suited to the empirical application, due to its differentiability and its output range of −1 to 1, which helps with the normalization of the outputs.

3.4.2 Interpreting the Probabilistic Latent Factors Generating Digital Twins

For the latent parameter layer of the digital twin architecture, which includes the latent variables and their prior distributions, recall that i indexes customer identifiers from 1 to N; k indexes firms from 1 to K; t indexes time from 1 to T. Customer-side factors include the following latent variables:

- $z_{ikt} \sim \mathcal{N}(\mu_{ikt}, \nu_{ikt})$: this latent factor has a Normal prior distribution. Because μ_{ikt} and ν_{ikt} are functions of an amortized neural network, this prior is highly flexible and encodes a wide range of customer characteristics, automatically accounting for interactions and nonlinearities. This parameter is interpreted as the *customer-side brand affinity factor*; it represents, for a given customer at a given time, their affinity with brand k.

- $\alpha_{jkt} \sim \mathcal{N}(0,1)$: The parameters α_{jkt} represent the *baseline* for question j for firm k at time t. It has a standard Normal prior distribution for simplicity.
- $\beta_{jl} \sim \mathcal{N}^{+}(0,1)$ The parameter β_j is interpreted as the *polarization* of question j in the l-th dimension of service quality, that is, how much question j elicits a response on the l-th service characteristic.

The firm-side factors include the following latent variables:

- $\phi_{ktl} \sim \mathcal{G}(\gamma_{ktl}, \omega_{ktl})$: The parameterization of a prior on ϕ_{ktl}, a firm-side latent factor on dimension l for firm k at time t, has a Gamma prior distribution. Because γ_{ktl} and ν_{ktl} are functions of an amortized neural network, this prior also is highly flexible; it encodes a wide range of firm characteristics, automatically accounting for interactions and nonlinearities.

Support for both ϕ_{ktl} and β_{jt} is the real positive line, for identification. The sign of z_{ikt} becomes then immediately interpretable, as explained in Subsection 4.3.

3.4.3 Model Layer for Ordered Categorical Outcomes

Because the target variables are all ordered categorical variables, I use an ordered logit specification. Let y^*_{ijkt} denote the latent response of respondent i to the entire set of J questions. Questions have different numbers of scale points: some questions have five scale points (0-4) whereas others have 10. For $M+1$ common and ordered cut points $\{c_m : c_{m-1} \leq c_m, m = 1, \ldots, M\}$ where $c_0 = -\infty$ and $c_M = +\infty$, latent utility values y^*_{ijkt} depend linearly on α_{jkt}, which are baseline values for question j at firm k; the customer-side brand affinity z_{ikt}; the polarization of question j in firm-side latent dimension l, β_{jl}; and the firm-side factors ϕ_{ktl}:

$$y^*_{ijkt} = \alpha_{jkt} + z_{ikt} \sum_{l=1}^{L} \left(\beta_j \phi_{kt}\right)_l + \varepsilon_{ijkt} \qquad \text{where } \varepsilon_{ijkt} \underset{i.i.d.}{\sim} EV(0,1) \tag{3.8}$$

The individual responses y_{ijkt} for a customer $i = 1, \ldots, N$ of firm $k = 1, \ldots, K$ at question $j = 1, \ldots, J_{targets}$, at time $t = 1, \ldots, T$ take the following values: $m = 1, 2, \ldots, M$ where M is the maximum number of scale points for question j. Because questions differ in their total number of scale points, m and M should have a subscript j, but I omit it for simplicity. The following holds:

$$y_{ijkt} = m \qquad \text{if } c_{j,m-1} \leq y^*_{ijkt} \leq c_{j,m} \tag{3.9}$$

where a Dirichlet prior model applies to ordinal probabilities, which serves to induce cut points indirectly. This approach enables a proper, principled prior on the cut points, which is useful when some categories are not strongly separated due to their data sparsity in some categories (Betancourt, 2020).

By marginalizing out the latent utilities y^*_{ijkt}, it is possible to write the probability of observing category m for question j in customer i of firm k at time t:

$$p(y_{ijkt}|c_{j,1},\ldots,c_{j,M}) = \begin{cases} \Pi(c_{j,1} - \alpha_{jkt} - z_{ikt} \sum_{l=1}^{L} \left(\beta_j \phi_{kt}\right)_l) & \text{if } m = 1 \\ \Pi(c_{j,m} - \alpha_{jkt} - z_{ikt} \sum_{l=1}^{L} \left(\beta_j \phi_{kt}\right)_l) & \\ -\Pi(c_{j,m-1} - \alpha_{jkt} - z_{ikt} \sum_{l=1}^{L} \left(\beta_j \phi_{kt}\right)_l)) & \text{if } 1 < m < M \\ 1 - \Pi(c_{j,m-1} - \alpha_{jkt} - z_{ikt} \sum_{l=1}^{L} \left(\beta_j \phi_{kt}\right)_l)) & \text{if } m = M \end{cases} \tag{3.10}$$

where $\Pi(.)$ is the cumulative distribution function of the Type I extreme value distribution, that is, the logistic function.

3.4.4 Inference and Implementation

The set of latent variables to infer is $\tilde{\mathbf{z}} = [\boldsymbol{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\omega}]$. The set of parameters to learn is $\boldsymbol{\theta} = \left[\boldsymbol{\theta}_f, \boldsymbol{\theta}_g\right]$. By writing $\mathbf{y}$ as the set of all observations (survey data and KPI), it is possible to approximate the posterior distribution $p_{\boldsymbol{\theta}}(\tilde{\mathbf{z}}|\mathbf{y})$.

Because of the size of the data, and the use of neural networks to parameterize the latent variables, exact inference (e.g., Markov chain Monte Carlo algorithms) is not feasible. Therefore, it is necessary to rely on approximate Bayesian inference; stochastic variational inference (SVI) aims at determining a variational distribution $q_{\boldsymbol{\lambda}}(\tilde{\mathbf{z}})$ that is as close as possible to the posterior $p_{\boldsymbol{\theta}}(\tilde{\mathbf{z}}|\mathbf{y})$ as measured by Kullback-Leibler (KL) divergence. Minimizing the KL divergence is equivalent to maximizing the evidence lower bound (ELBO) on the log marginal probability of the data $\log p_\theta(\mathbf{y})$, with $\log p_\theta(\mathbf{y}) \geq \text{ELBO}$ and $\log p_\theta(\mathbf{y}) - \text{ELBO} = \text{KL}\left(q_{\boldsymbol{\lambda}}(\tilde{\mathbf{z}})\right) \| p_{\boldsymbol{\theta}}(\tilde{\mathbf{z}}|\mathbf{y})\right)$.

The evidence lower bound (ELBO) is:

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q_{\boldsymbol{\lambda}}(\tilde{\mathbf{z}})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{y}, \tilde{\mathbf{z}})\right] - \mathbb{E}_{q_{\boldsymbol{\lambda}}(\tilde{\mathbf{z}})}\left[\log q_{\boldsymbol{\lambda}}(\tilde{\mathbf{z}})\right] \tag{3.11}$$

The ELBO creates two expectations with respect to the variational distribution. The first expectation, $\mathbb{E}_{q_{\boldsymbol{\lambda}}(\tilde{\mathbf{z}})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{y}, \tilde{\mathbf{z}})\right]$, represents the expected log-likelihood of the data given the model parameters, which encourages densities that place their mass on configurations of the latent variables that explain the observed data (Blei, Kucukelbir, and McAuliffe, 2017). The second expectation, $\mathbb{E}_{q_{\boldsymbol{\lambda}}(\tilde{\mathbf{z}})}\left[\log q_{\boldsymbol{\lambda}}(\tilde{\mathbf{z}})\right]$, is the negative divergence between the variational density and the prior. Maximizing the ELBO is akin to finding a balance between encouraging the model to fit the data well (maximizing the first term) and encouraging densities close to the prior (maximizing the second term).

In line with standard practice, this study uses mean-field variational approximation. The model implementations relies on the machine learning framework Google

JAX for fast computation on Graphics Processing Unit (GPU), and the NumPyro probabilistic programming language (Phan, Pradhan, and Jankowiak, 2019). With the Adam optimization algorithm (Kingma and Ba, 2014), a Monte Carlo version of the loss function is optimized in Equation (3.11) and a test set is used to determine all model hyperparameters, namely, the number of hidden layers per neural network, number of hidden units per neural network, and number L of latent firm-side dimensions.

3.4.5 Digital Twin Generative Process

To summarize, the specification is such that for all $m = 1, \ldots, M$, $i = 1, \ldots, N$, $j = 1, \ldots, J$, $k = 1, \ldots, K$ and $t = 1, \ldots, T$:

$$
\begin{aligned}
&\begin{bmatrix}\mu_{ikt}\\ \nu_{ikt}\end{bmatrix} = f(\mathbf{x}_{it}, \boldsymbol{\theta}_f) && \text{where } f \text{ is a feed-forward neural network}\\
&\begin{bmatrix}\gamma_{kt}\\ \omega_{kt}\end{bmatrix} = g(\mathbf{x}_{kt}^{(\mathrm{KPI})}, \boldsymbol{\theta}_g) && \text{where } g \text{ is a feed-forward neural network}\\
&\beta_j \sim \mathcal{N}^+(0,1)\\
&z_{ikt} \sim \mathcal{N}(\mu_{ikt}, \nu_{ikt})\\
&\phi_{ktl} \sim \mathcal{G}amma(\gamma_{ktl}, \omega_{ktl})\\
&y^*_{ijkt} = \alpha_{jkt} + z_{ikt}\sum_{l=1}^{L}\left(\beta_j\phi_{kt}\right)_l + \varepsilon_{ijkt} && \text{where } \varepsilon_{ijkt} \underset{i.i.d.}{\sim} EV(0,1)\\
&\pi(\mathbf{c}_j|\kappa,\lambda) = \mathcal{D}(p(\mathbf{c}_j,\lambda)|\kappa).|J(\mathbf{c}_j,\lambda)|\\
&y_{ijkt} = m && \text{if } c_{j,m-1} \leq y^*_{ijkt} \leq c_{j,m}
\end{aligned}
$$

where $\mathcal{D}$ is the Dirichlet probability density function, and J is the Jacobian matrix. A uniform Dirichlet prior $\kappa = (1, 1 \ldots, 1)$ and the anchor point $\lambda = 0$ identify the model, without loss of generality. Figure 3.2 provides an illustration of the full digital twin architecture. The hybrid deep learning and probabilistic generative framework allow the best of both worlds: high flexibility and representation power of the amortized neural network on the left hand side of the illustration, in orange, and high interpretability and theory-based for counterfactual reasoning on the right hand side, in green.

3.4.6 Identification

Two core empirical challenges prevent marketing analysts from drawing respondent-level counterfactual inferences from the observed outcomes of brand surveys, or even when utilizing discriminative models. Among the T repeated cross-sectional surveys for a carrier k, it is highly unlikely that any respondent would repeatedly manifest across surveys. This data regime is not only common in commercial brand

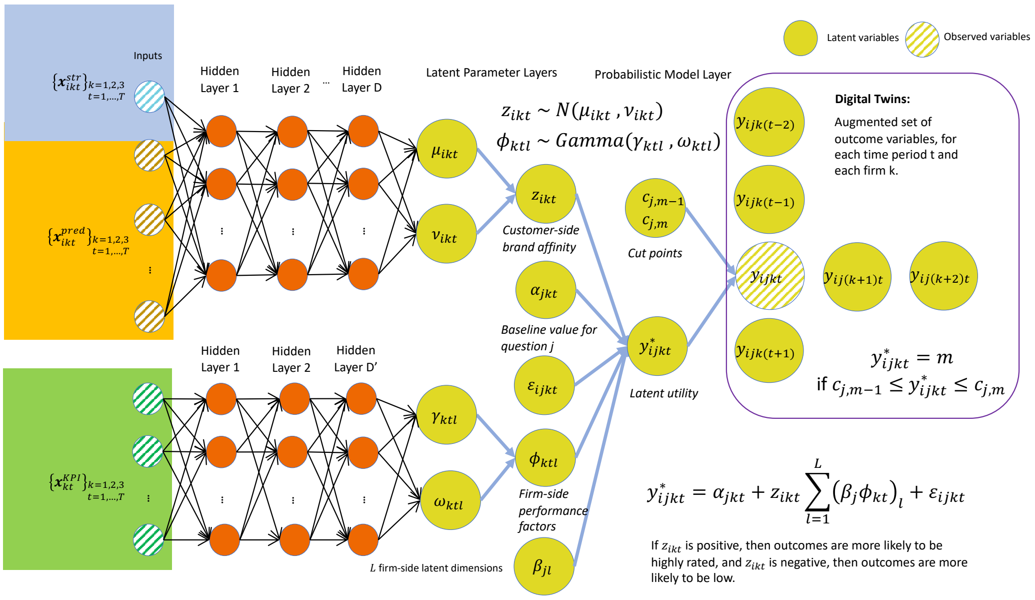

Latent variables
Observed variables
Inputs
$\{x_{ikt}^{str}\}_{k=1,2,3}^{t=1,\ldots,T}$
$\{x_{ikt}^{pred}\}_{k=1,2,3}^{t=1,\ldots,T}$
$\{x_{kt}^{KPI}\}_{k=1,2,3}^{t=1,\ldots,T}$
Hidden Layer 1
Hidden Layer 2
Hidden Layer D
Hidden Layer 1
Hidden Layer 2
Hidden Layer D'
Latent Parameter Layers
Probabilistic Model Layer
$z_{ikt} \sim N(\mu_{ikt}, \nu_{ikt})$
$\phi_{ktl} \sim Gamma(\gamma_{ktl}, \omega_{ktl})$
μ_{ikt}
ν_{ikt}
z_{ikt}
Customer-side brand affinity
α_{jkt}
Baseline value for question j
ε_{ijkt}
γ_{ktl}
ω_{ktl}
ϕ_{ktl}
Firm-side performance factors
β_{jl}
L firm-side latent dimensions
$c_{j,m-1}$
$c_{j,m}$
Cut points
y_{ijkt}^*
Latent utility
Digital Twins:
Augmented set of outcome variables, for each time period t and each firm k.
$y_{ijk(t-2)}$
$y_{ijk(t-1)}$
y_{ijkt}
$y_{ij(k+1)t}$
$y_{ij(k+2)t}$
$y_{ijk(t+1)}$
$y_{ijkt}^* = m$
if $c_{j,m-1} \leq y_{ijkt}^* \leq c_{j,m}$
$y_{ijkt}^* = \alpha_{jkt} + z_{ikt} \sum_{l=1}^{L} (\beta_j \phi_{kt})_l + \varepsilon_{ijkt}$
If z_{ikt} is positive, then outcomes are more likely to be highly rated, and z_{ikt} is negative, then outcomes are more likely to be low.

FIGURE 3.2: A deep probabilistic architecture of the modeling framework. An amortized inference neural networks (in orange, top) take survey data question as inputs and parameterize latent customer-side brand affinity z_{ikt}. Another amortized inference neural network (in orange, bottom) takes KPI data as input and parameterize firm-side performance factors ϕ_{ktl}. These two types of latent variables are then combined in the probabilistic model layer according to the Equation (3.8) as a latent customer-level utility. This latent utility is then evaluated against latent question-level cut points to present ordered categorical variables for target customer satisfaction questions.

surveys, but across typical survey designs in the social sciences (Groves et al., 2011). Hence, the first challenge is that such pseudo-longitudinal setting disallows the application of standard longitudinal panel models to analyze these types of data. Second, except in rare cases, the presence of any customer i of carrier k in the U.S. wireless telecom sector precludes the possibility of their simultaneous presence as a customer of competing carrier k'. These two identification challenges motivate the development of Digital Marketing Twins.

In this section, I establish the mechanism of Digital Marketing Twins' generative framework in identifying respondent-level counterfactual outcomes across K carriers and T periods, while assuming a data generating process of the survey outcomes whereby any respondent i is only ever observed for a single carrier k, in a single period t.

The counterfactual identification strategy described below extends missing data approaches found in marketing and statistics (Rubin, 1976; Little and Rubin, 2019; Rossi and Allenby, 2003) to formalize a set of antecedent modeling assumptions and how they translate into the consequent posterior predictive distribution. Where appropriate, testable vs. untestable assumptions necessary for internal validity are delineated, along with their implications on external validity (i.e., managerial actionability). Finally, I draw parallels to the closely related missing data paradigm of the Rubin Potential Outcomes framework (Rubin, 1974), as well as contrast the counterfactuals under Digital Marketing Twins versus causal estimands based on potential outcomes – namely, in the analysis of brand surveys, there does not exist the notion of treatment interventions, which is the focal design of any causal inference undertaking.

Target counterfactual. The Digital Marketing Twins framework enables the identification of the posterior predictive distribution of:

$$p^{\grave{}}(\boldsymbol{z}_{ik'}|\mathbf{x}_{ikt}) \tag{3.12}$$

where $k' \neq k$ such that $\boldsymbol{x}_{ikt}$ is the J-length vector of observed survey outcomes for customer i of firm k who responded in period t, $\mathbf{z}_{ik'}$ is a T-by-D matrix of customer-side brand-affinity if i were – counterfactually – a customer of firm k' instead.

Let $\mathcal{D}$ denote the observed data generating process, which face the two aforementioned repeat cross-sectional empirical limitations; and $\mathcal{D}^*$ denote an oracle data generating process whereby all N respondents appear in all T periods and for all K carriers. Samples drawn from the above target distribution (Eq. 3.12) are defined as counterfactual samples if the corresponding stationary posterior given $\mathcal{D}$ is identical to the stationary posterior given $\mathcal{D}^*$.

If given oracle data with sufficiently large N, $\{\boldsymbol{x}\}_{i=1}^{N} \subseteq \mathcal{D}$, an arbitrarily flexible generative model can robustly and consistently infer the above target distribution (Eq. 3.12) by simply learning a bijective mapping of a respondent's outcomes from any period t, under any carrier k, to respondent's own outcomes for any other period $t' \in \{1, \ldots, T\}$ and carrier $k' \in \{1, \ldots, K\}$. However, oracle data are infeasible to collect, both due to cost of carrying out large-scale longitudinal brand surveys as well as the market reality that the vast majority of U.S. wireless consumers procure service from a single carrier at any time. Therefore, the identification of the counterfactual posterior predictive distribution from observed data relies on (1) desired empirical regularities in $\mathcal{D}^*$ that have equivalents in $\mathcal{D}$, and (2) undesired empirical regularities in $\mathcal{D}$ that must be controlled for via model specifications.

Formally, the posterior predictive distribution, conditional on observed data, that produces the desired counterfactuals must meet the following criterion on the KL-divergence, a measurement of the difference between distributions:

$$D_{KL}\left\{p^{\grave{}}(\mathbf{z}_{ik'}|\mathbf{x}_{ikt} \subset \mathcal{D}^*) \| p^{\grave{}}(\mathbf{z}_{ik'}|\mathbf{x}_{ikt} \subset \mathcal{D})\right\} = 0 \tag{3.13}$$

- **Assumption 1: Ignorability in y.** Extending the classic econometric age-period-cohort (APC) approach (e.g., Mason et al., 1973; Yang, 2006) to modeling repeat cross-sectional panels via partial pooling, here ignorability posits that idiosyncratic differences across time and carriers can be deconfounded (i.e., ignorable) via the latent variables α_{jkt} and ϕ_{kt}. Whereas the APC framework assumes all cohort differences (i.e., selection artifacts and other unobservables) are captured by the additive parameter α_{jkt}, in Digital Marketing Twins, this deconfounding mechanism is extended to also include the multiplicative term ϕ_{kt}. Together, as amortized parameters, α_{jkt} and ϕ_{kt} serve to flexibly control for confounding arising from unobservable factors that would bias the counterfactual inference of $\mathbf{z}_{ik'}$ in repeat cross-sectional settings.

- **Assumption 2: Comparability in** x**.** As shown in the model-free evidence, the brand surveys exhibit strong overlapping empirical support in input features x across periods and carriers. Having this overlap in the observed data generating process $\mathcal{D}$ signifies that – despite any respondent i is only ever observed for a single carrier k, in a single period t – the distributions of x are comparable across any other period $t' \in \{1, \ldots, T\}$ and carrier $k \in \{1, \ldots, K\}$. Under comparability, any sample from the posterior, when conditioned on identical values of x but varying in period and/or carrier, can be considered as interpolations within the empirical support – i.e., robust and consistent to the equivalent posterior under $\mathcal{D}^*$.

- **Assumption 3: Exchangeability in** z**.** Given assumptions 1 and 2, it follows that $p_{\grave{}}(\mathbf{z}_{ik'}|\mathbf{x}_{ikt})$ (Eq. 3.12) is robust and consistent to any permutation in the indexing of z and x. Should the indexing entail $p_{\grave{}}(\mathbf{z}_{ik'}|\mathbf{x}_{ikt})$, then we can interpret this posterior predictive distribution as the counterfactual distribution of the customer-side brand-affinity of customer i if they were – exchangeably – a customer of firm k' instead.

Lastly, while Eq. 3.12 has a canonical form of a conditional distribution, its validity is asserted through exchangeability, which is a weaker assumption than conditional independence. Whereas the latter can be assessed through empirical hypothesis testing, the former arises in counterfactual and missing data contexts where the validity of the inference on the unobserved outcome(s) must arise from assumptions on the data generating process, as done above. In summary, recognizing the “chasm" between the observed data generating process of surveys, $\mathcal{D}$, versus the ideal data generating process $\mathcal{D}^*$, the Digital Marketing Twins framework utilizes flexibly parameterizations to control for observed and unobserved confounders (Assumption 1), as well as exploits essential empirical regularities in $\mathcal{D}$ that mimics $\mathcal{D}^*$ (Assumption 2), to establish that the counterfactual inferences capable of being drawn from Eq. 3.12 are *exchangeably* valid across time and firms (Assumption 3) – despite the absence of the ideal, but unrealistic, longitudinal survey data.

3.4.7 Relation to Prior Literature

Variational Autoencoders

My model is novel in its use of customer-level predictors and strategic variables that parameterize an amortized neural network for high flexibility, that output structured latent variables that can be subsequently interpreted by marketers, and generate a coherent model of customer satisfaction. However, neither the use of amortized neural networks nor the use of Bayesian optimization in marketing is novel.

The model resembles a Variational Autoencoder (VAE) (Kingma and Welling, 2013), which is also a generative model that also uses variational inference for learning. The differences lie primarily in the specific structure of the model and the form of the decoder. In the VAE, the encoder is a neural network that takes the observed data as inputs, then outputs parameters of a distribution over the latent variables. The proposed model has two such "encoders", f and g, each of which produces parameters for different distributions over subsets of the latent variables, such that f encodes μ_{ikt} and ν_{ikt} for customer-side brand affinity z_{ikt}, and g encodes γ_{kt} and ω_{kt} for firm-side factors ϕ_{kt}.

In a VAE, the latent variables capture unobserved factors of variation in the data, whereas z_{ikt} and ϕ_{kt} capture observed factors of variation in the data, because they are parameterized by survey and KPI inputs. The only unobserved factors of variations come through the type I extreme value that affects latent utilities.

The decoder in a VAE takes the latent variables and generates parameters for the distribution over the observed data. The equations involving y_{ijkt} and y^*_{ijkt} can be interpreted as part of a kind of decoder that uses the latent variables, together with an ordered logit model, gives a distribution over the observed variable y_{ijkt}. However, unlike a VAE, this decoder does not involve a neural network but is determined by an ordered logit model and a latent factor model that decomposes firm-side and customer-side effects.

Bayesian Models in Political Science

The proposed model connects loosely with political science literature, through the notion of Bayesian ideal points.[5] In political science, a latent factor model quantifies lawmakers' political preferences using roll-call votes (Jackman, 2001; Clinton, Jackman, and Rivers, 2004). Lawmakers yay or nay voters on a shared set of bills can be coded as a binary variable y_{ij} for lawmaker i voting for bill j. Each lawmaker is assumed to have a latent variable z_i, also known as the ideal point, and per-bill latent variables (α_j, β_j). The vote can be modeled as a Bernoulli distribution

[5] Note that the ideal point in political science has a different meaning that in marketing; in marketing, it refers to the hypothetical product attributes or characteristics that a consumer would perceive as perfect, indicating their absolute preference. Market research and product development often use the ideal point concept to tailor offerings to align closely with consumer desires, increase product appeal, and ensure market success.

$y_{ij} \sim \mathcal{N}(\sigma(\alpha_j + z_i\beta_j))$ where $\sigma(x) = \frac{1}{1+\exp(-x)}$. The customer-side brand affinity functions as an ideal point, such that each customer's measurement on target variables is akin to rating each company.

3.5 Model Results

3.5.1 Fit and Benchmarks

The comparison of the proposed model with three benchmarks confirms its validity, as the goodness-of-fit metrics in Table 3.3 reveal. The three competing specifications are as follows: Model (1) is a linear version of the proposed model, in which the neural networks have been replaced by a linear layer. It is akin to a traditional SEM, with observed inputs and outputs, but has latent variables parameterizing the relationship between inputs and outputs. A traditional SEM is therefore a special case. Model (2) omits individual predictive and strategic variables from the full specification, but retains individual level customer-side brand affinity z_{ikt} for all $i = 1, .., N$, $k = 1, ..., K$, $t = 1, ..., T$ as "random effects". Model (3) omits KPI, but retains firm level performance factors ϕ_{ktl} for all $l = 1, ..., L$, $k = 1, ..., K$, $t = 1, ..., T$ as "random effects".

Figure 3.3 shows evidence of convergence, in that the average training and test loss (i.e., negative evidence lower bound – ELBO) decrease rapidly before stabilizing. The average test loss is close to the training loss, suggesting good generalizability. The greater variance of the test loss results from the test sample being smaller than the training sample.

In terms of goodness-of-fit, the proposed model (Model (4)) consistently performs better than other models across all carriers and metrics. The average training and test losses are lowest in this model, indicating that it offers the best fit to the data. For example, the Average Training Loss for Model (4) is 11.54, lower than the corresponding values for other models. The same trend is apparent in the Average Test Loss, in which Model (4) outperforms other models with a loss of 12.81. Furthermore, with regard to the Mean Absolute Error (MAE) for each carrier across different tasks, Model (4) generally exhibits the smallest error, suggesting it the most capable of accurately reconstructing the data. Some exceptions involve the Phone LTR and Carrier Satisfaction for Carriers 1 and 2, and the Retention Likelihood for Carrier 3, for which Model (4) does not perform best. However, the overall performance of Model (4) remains superior.

Thus, Table 3.3 suggests that a neural network model that includes individual predictors and KPI performs best among the presented models across a variety of reconstruction tasks. It also indicates the complexity of the relationships in the data and the ability of neural networks to better capture these complex relationships and extract predictive latent features.

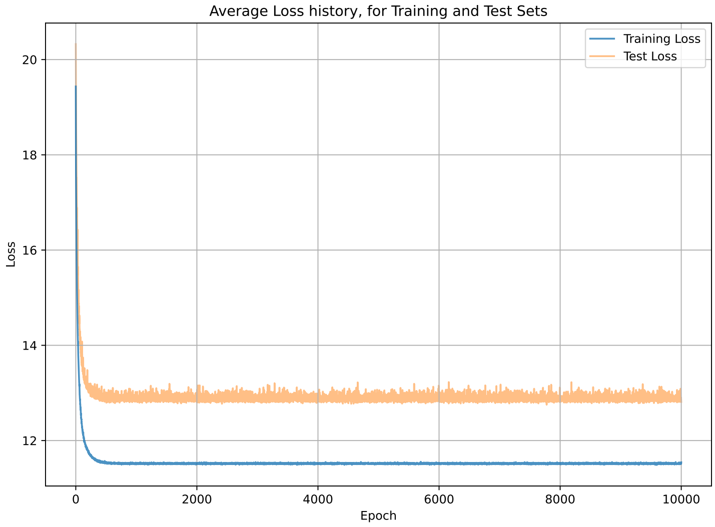


FIGURE 3.3: Average training and test loss (ELBO), over all individuals in training and test sets, respectively.
Notes: The plot suggests that convergence is reached after about 1000 epochs, but the model was trained for 10,000 epochs in total. Test loss is slightly greater than training loss, but remains constant after convergence, as expected.

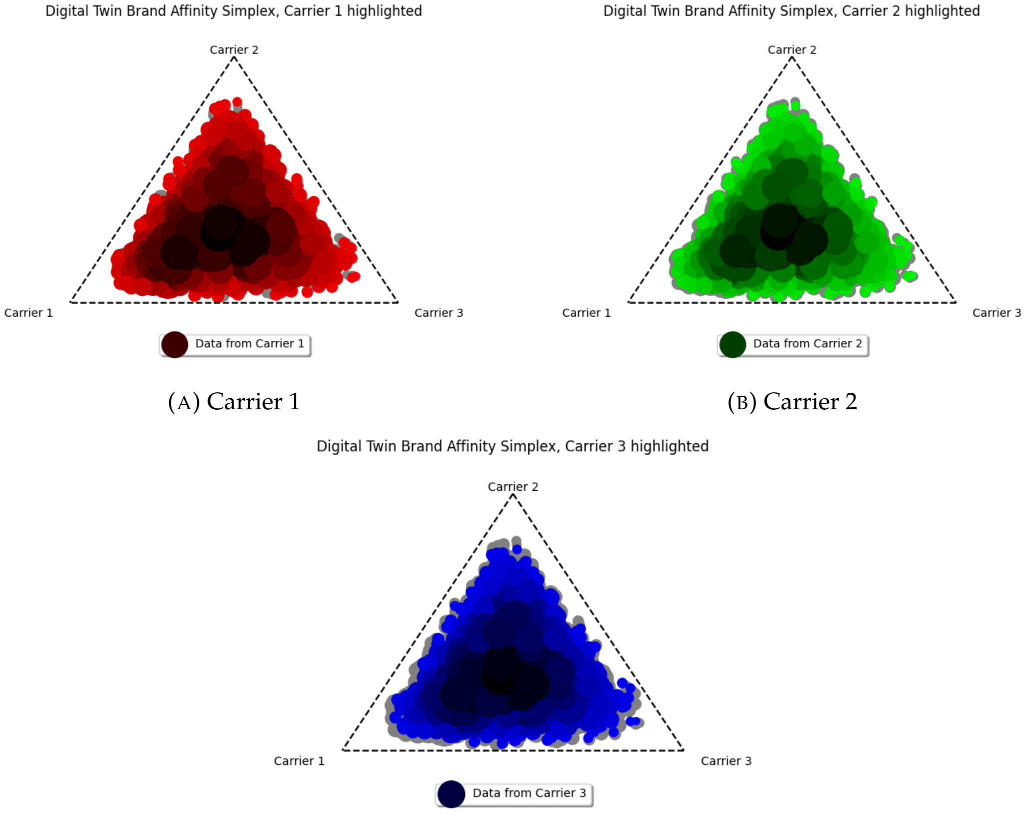
3.5.2 Analyzing and Interpreting the Probabilistic Latent Factors

A crucial aspect of the proposed framework is its ability to analyze and interpret the estimated probabilistic latent factors while relaxing the functional form between various predictors and these factors for maximum flexibility.

Figure 3.4 plots the counterfactual customer-side brand affinities z^*_{ikt} using posterior means across all customers. A darker color signifies higher density. Each dot represents the triplet $(z_{i1t}, z_{i2t}, z_{i3t})$ where $t = 0$, after transformation through a softmax function, which is then converted in a barycentric coordinate system to obtain values that live in a three-dimensional simplex. For example, a customer in Carrier 2 with a dot near the "Carrier 1" vertex reflects that the brand affinity level that customer would obtain if they were assigned to Carrier 1, i.e., the *digital twin* of that customer under the counterfactual regime that this customer's carrier is now Carrier 1.

One interesting phenomenon to notice is the higher density of customers toward Carrier 1, for all carriers' customer bases. This digital twin representation suggests a large group of customers who would have a high brand affinity with Carrier 1, if they were ever assigned to be its customers. This latent asymmetry in brand affinity could not be identified without a rigorous counterfactual analytical framework. Carrier 1 likely should target this group of prospective customers to steal them from Carrier 2's and Carrier'3 customer bases.

Figure 3.5 plots baseline values α_{jkt} for each target variable over time, showing



(A) Carrier 1

(B) Carrier 2

(C) Carrier 3

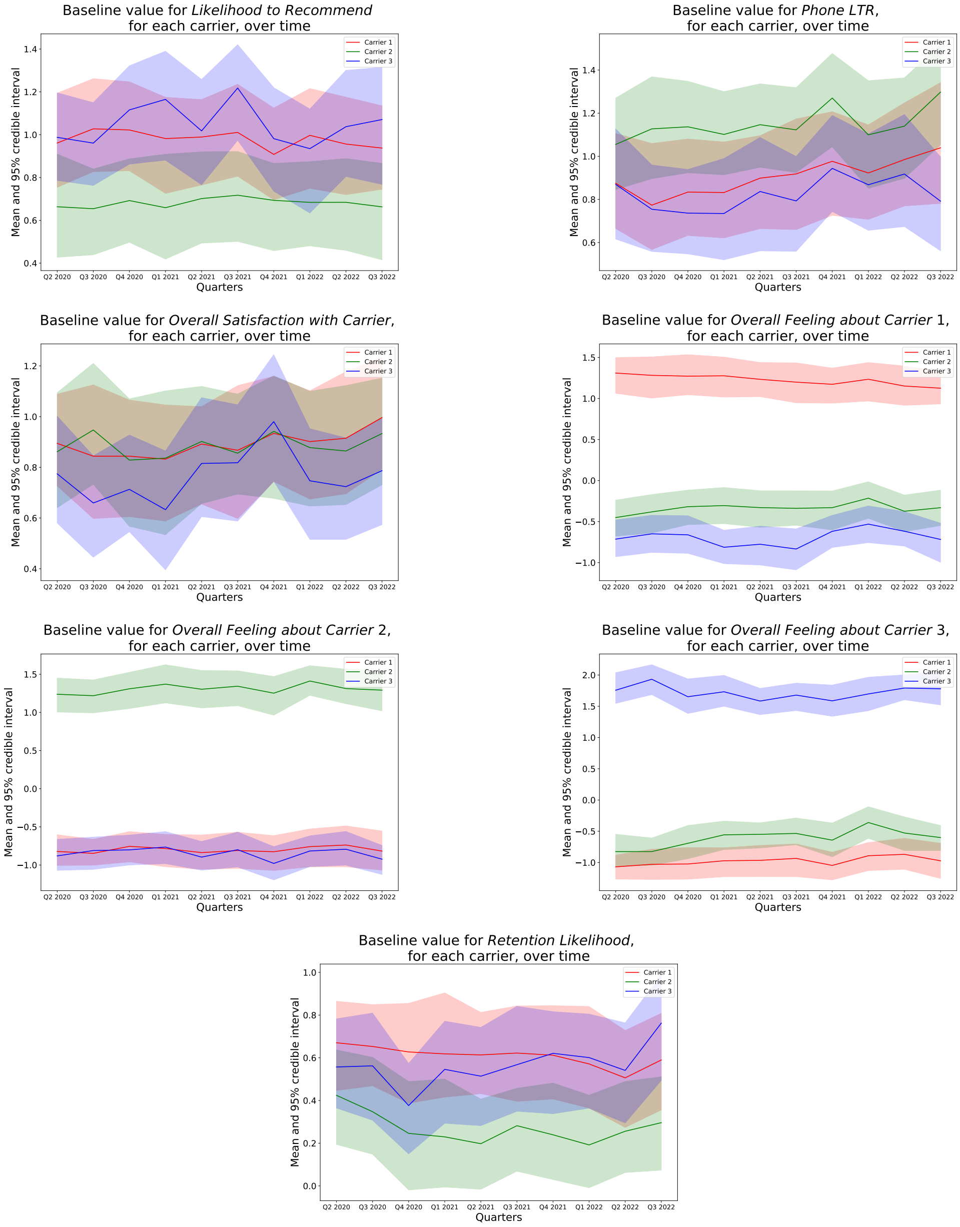
FIGURE 3.4: Plotting counterfactual customer-side brand affinities z^*_{ikt}, using posterior means across all customers.
Notes: A higher density is signified by darker colors. Brand affinities have been transformed using a softmax to fit into a simplex. Each dot represents a customer from a given carrier, and can be projected onto the edges of the triangle to reveal the manifest *digital twins*, i.e., counterfactual brand affinities summarizing target variables.

Models		(1)	(2)	(3)	(4)
ϕ		Linear	NN	w/o KPI	NN
z		Linear	w/o indiv. predictors	NN	NN
No. epochs		10000	10000	10000	10000
Avg. Training Loss		12.22	14.06	11.57	**11.54**
Avg. Test Loss		13.49	15.81	12.94	**12.81**
Test Mean Absolute Error (MAE)					
Carrier 1	LTR	0.84	1.59	0.67	0.64
	Phone LTR	1.39	1.57	1.24	1.27
	Carrier Satisfaction	1.10	1.51	0.95	1.00
	Overall Feeling Carrier 1	0.53	0.71	0.49	0.42
	Overall Feeling Carrier 2	0.68	0.61	0.68	0.64
	Overall Feeling Carrier 3	0.55	0.47	0.56	0.50
	Retention Likelihood	0.74	0.90	0.69	0.71
Carrier 2	LTR	0.88	1.75	0.69	0.67
	Phone LTR	1.31	1.48	1.15	1.17
	Carrier Satisfaction	1.11	1.53	0.97	1.01
	Overall Feeling Carrier 1	0.70	0.70	0.69	0.68
	Overall Feeling Carrier 2	0.51	0.75	0.47	0.43
	Overall Feeling Carrier 3	0.61	0.57	0.60	0.57
	Retention Likelihood	0.85	1.01	0.81	0.82
Carrier 3	LTR	0.84	1.54	0.63	0.62
	Phone LTR	1.35	1.60	1.20	1.21
	Carrier Satisfaction	1.11	1.54	0.95	1.00
	Overall Feeling Carrier 1	0.75	0.69	0.75	0.70
	Overall Feeling Carrier 2	0.76	0.70	0.77	0.71
	Overall Feeling Carrier 3	0.49	0.68	0.47	0.38
	Retention Likelihood	0.74	0.89	0.68	0.69

TABLE 3.3: Goodness-of-fit metrics. Training and Testing Loss, and Mean Absolute Error for Reconstruction Tasks. Model (4) is benchmarked against nested versions (1,2,3). Model (1) assumes a linear link between inputs (predictive variables, strategic variables, key performance indicators) and latent variables. Model (2) omits individual predictors. Model (3) omits key performance indicators.

posterior mean and 95% credible intervals. These values represent the base utility for a given question, at a given time, and for a given carrier, after accounting for non-parametric variations in individual predictors or KPI. Baseline values for LTR seem to be lower for Carrier 2, though they seem higher in terms of recommending their carrier's phone. Unsurprisingly, baseline values for Overall Feeling about Carriers 1, 2, and 3 are higher for corresponding customer bases. Finally, baseline values for Retention Likelihood for Carrier 1 and 3 are higher than for Carrier 2; customers of Carrier 2 are more likely to switch to the competition, on average.

Figure 3.6 plots the carrier performance loading on each target variable over time. The evidence is more mixed, because reflecting considerable uncertainty, as also indicated by the credible interval. This uncertainty is propagated into the model's predictive performance, explaining why the model with KPI performs only marginally better than the model without them. An interesting aspect to notice is that carrier



Baseline value for *Likelihood to Recommend* for each carrier, over time

Baseline value for *Phone LTR*, for each carrier, over time

Baseline value for *Overall Satisfaction with Carrier*, for each carrier, over time

Baseline value for *Overall Feeling about Carrier* 1, for each carrier, over time

Baseline value for *Overall Feeling about Carrier* 2, for each carrier, over time

Baseline value for *Overall Feeling about Carrier* 3, for each carrier, over time

Baseline value for *Retention Likelihood*, for each carrier, over time

FIGURE 3.5: Plot of the baseline values α_{jkt} for each target variable over time.
Notes: These values represent the base utility for a given question, at a given time, and for a given carrier, abstracting away from individual predictors or key performance indicators.

performance loading on Overall Feeling about Own Carrier and Competition's Carrier are well informed, as is also reflected in the lower test MAE shown above in Model (4) compared with Model (3).

3.5.3 Application: Personalized Marketing Communication Campaigns using Grid Search on Strategic Variables

After the model is trained, digital twins can be used to optimize personalized marketing communications campaigns. Certain features of telecommunications services are complex and not well understood or known by customers. The proposed framework allows marketers to automatically rearrange current communication strategies automatically, to focus on the most critical aspects of customer satisfaction, at the individual level.

With a grid search on strategic variables for all customers in the test sample of the dataset, I decrease each customer's current value by 2 points for each strategic variable with at least five point scales, then gradually increase each current value by 0.1 increments, until it reaches +2 points. These changes are simultaneously implemented across all customers in 2020 Q2. Figure 3.7 plots individual-level responses in brand affinity after changing strategic variables from -2 to 2, using the posterior mean for z_{ikt} as a summary statistic. Carrier 2 seems to have a group of customers with lower brand affinity. Carriers 1 and 3 are remarkably close in terms of brand affinity level. The plots also offer strong evidence of customer-level nonlinearities for two strategic variables. A more positive response by customers to a change in wireless service within the next six months would induce a lower brand affinity, up to a certain point. However, a more positive response by customers to a satisfaction question on their providers' network speed would induce a rapid increase in brand affinity for most customers.

3.6 Optimizing Customer Satisfaction with Bayesian Optimization

The Digital Marketing Twins model, presented in Section 4, estimates a probabilistic generative model of customer-side brand affinity and firm-side performance factors. The analysis in Section 5.3 showed how posterior inference on the model can be used by managers to undertake campaign personalization, which fits into the broader category of discriminative, or segmentation, based analysis in marketing research. The assumption of partial equilibria in customers' preferences and perception of brands and their offerings, conditional on firm-side (e.g., marketing) efforts, is fundamental to any applied discriminative marketing analysis; in the context of the proposed model's inference for the U.S. wireless telecom industry, it is equivalent to assuming

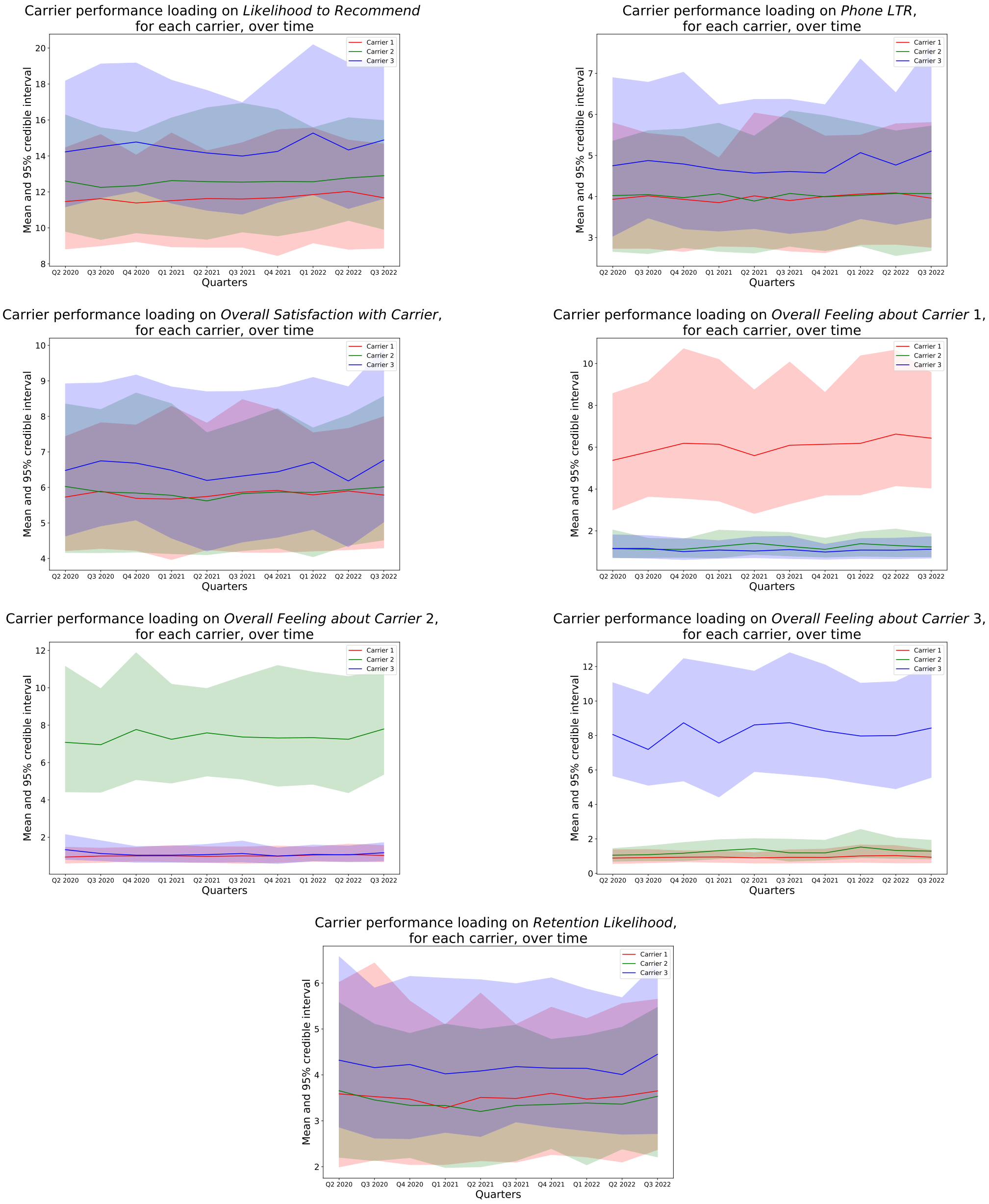
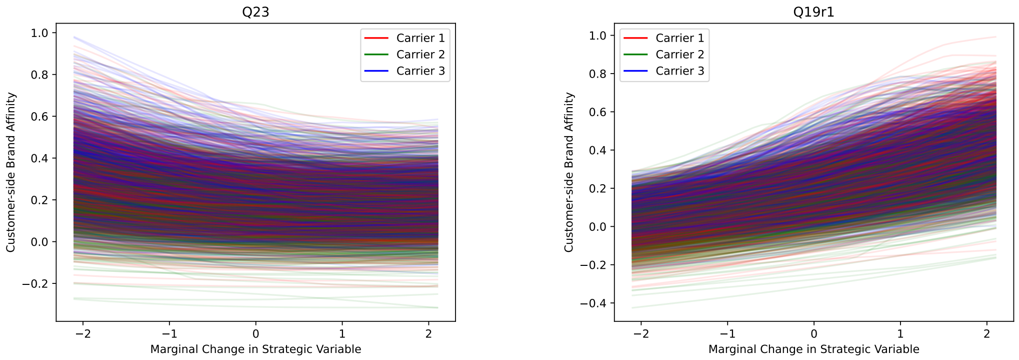


FIGURE 3.6: Plot of the carrier k performance loading on each target variable j ($\sum_{l=1}^{L} \beta_{jl}\phi_{lkt}$) over time t.
Notes: These values represent the contribution from KPI to target questions. These values are multiplied to brand affinity.



(A) "How likely are you to change anything about your current wireless service in the next 6 months?"

(B) How satisfied are you with your carrier's performance on the following aspect of your overall wireless service experience: Network Speed

FIGURE 3.7: Plotting variations in counterfactual customer-side brand affinities z^*_{ikt} in response to marginal changes in strategic variables, between -2 and 2. Each line represents an individual customer's response from a given carrier, in Q2 2020. These plots show evidence of individual-level nonlinearities in brand affinity responses to marginal changes in strategic variables.

that the distribution across the customer base's brand affinity is invariant, because of ongoing marketing strategies and realizations of capital expenditure.

This assumption is reflected in the *retroactive* and *tactical* nature of personalization campaigns, which, due to the current strengths and weaknesses across the dimensions of an individual's brand affinity, resulting from customers' experiences, usages, and interactions, should tailor forthcoming marketing communications to highlight strengths and ameliorate weaknesses. By leveraging the Digital Marketing Twins generative framework, this section highlights its potential as a *proactive* and *strategic* tool, grounded in the statistical perspective that counterfactual reasoning is a missing data problem, to rationalize and optimize a wireless telecom carrier's marketing strategy. Specific marketing strategy dimensions lead to a more optimal distribution of a customer base's brand affinity, by taking into consideration not only differences across customers, but also the competitive landscape.

From an optimization perspective, should the generative model be convex and smooth everywhere, then an optimal firm-level marketing policy – realized in this specific exercise as relative emphases across the 33 strategic dimensions – can be found through differential calculus on the model's functional form. However, two computational challenges arise that require the use of Bayesian optimization: (1) the fitted model specification is almost surely nonconvex, and (2) a brute-force search of the policy grid creates an infeasible $O(n^d)$ time complexity.[6]

[6] For example, a grid of 1.0, by 0.1 grid size, would translate into $10^3$3 calculations to be evaluated. State-of-art GPUs have 20,000 cores, each of which can (optimistically) evaluate the model to calculate the individual-level brand affinity of all customers in one millisecond. This search thus would still take 1.58 quintillion years to complete.

In dealing with these two computational challenges, Bayesian optimization is a powerful tool that can resolve black-box global optimization problems, particularly those with expensive function evaluations (Letham and Bakshy, 2019). Bayesian optimization combines principles of exploration and exploitation, and uses acquisition functions to navigate the optimizable design space effectively and feasibly. I introduce a novel expected improvement (EI) acquisition likelihood that will enable marketers to identify *paths of least resistance* to improve overall and individual-specific brand affinity within a competitive landscape.

My framework is novel in that it combines a generative model to use with Bayesian optimization for policy search (e.g., Athey, Wager, et al., 2017). The Digital Marketing Twins generative model can be viewed as a form of offline simulators (Letham and Bakshy, 2019), that elicits the effects of changes to a system more efficiently, which in this case is the brand affinity distribution of the customer base. Formally, as an offline simulator in a policy search framework, the Digital Marketing Twins framework allows (1) exploration of brand affinity distributions under different marketing policy configurations (2) computation of both individual- and aggregate- level "rewards" (i.e., increase in brand affinity relative to other brands), and (3) a counterfactual estimate of alternative policies on the same customer or segment (Bottou et al., 2013; Dudík et al., 2014).

The modeling of the counterfactual brand affinity response surface uses the popular Bayesian optimization implementation through a Gaussian process (GP) prior, denoted as:

$$z^*_{ikt} \sim \mathcal{GP}\left(m(.), K(.,.)\right) \tag{3.14}$$

I denote brand affinity evaluations from the GP as z^*_{ikt} to indicate that they are counterfactual outcomes that do not necessarily map onto z_{ikt} inferred from the data used to train the Digital Marketing Twins generative model. For this reason, the GP prior is seen as a distribution over function spaces, and analogously, the goal of Bayesian optimization is to learn a global policy response set and identify feasible optima. The prior is characterized by a mean function $m(\mathbf{x}^{(str)}) = \mathbb{E}[f(\mathbf{x}^{(str)})]$ and a covariance function $K(\mathbf{x}^{(str)}, \mathbf{x}'^{(str)}) = \text{cov}[f(\mathbf{x}^{(str)}), f(\mathbf{x}'^{(str)})]$. The covariance function specifies the covariance between any two points in the design space, which as noted, consists of thirty-three strategic variables. Specifically, I use the ARD-RBF kernel, hyperparameterized by τ (amplitude) and l_j (policy-dimension specific lengthscale). The advantages of the ARD-RBF kernel are that it undertakes variable selection, and it is infinitely differentiable, which enables me to capture complex nonlinear interactions through the policy space:

$$K(\mathbf{x}^{(str)}, \mathbf{x}'^{(str)}) = \tau^2 \exp\left(-\frac{1}{2}\sum_{j=1}^{J}\left(\frac{\mathbf{x}_j^{(str)} - \mathbf{x}_j'^{(str)}}{l_j}\right)^2\right) \tag{3.15}$$

In the current study context, the nonparametric GP prior is especially useful as a policy response surface model for Bayesian optimization. First, it provides uncertainty estimates for unobserved points, which are crucial for applying an explore/exploit algorithm. Second, the mean and variance predictions are available in closed form, enabling fast gradient optimization when identifying the next optimal point to test. Third, the smoothness assumption allows efficient exploration of the complex, nonlinear relationships between strategic and target variables from the surveys, which were learned using neural networks during training.

Finally, to choose policies for future evaluation, I use the Expected Improvement (EI) acquisition function that is integrated over the posterior distribution in brand affinity, and rationalized over explore/exploit dynamics in this context. The optimization problem I address aims to maximize an objective $\mathbb{E}[z^*]$, with the constraints $c_j(\mathbf{x}_j^{(str)}) \geq 0$ for $j = 1, ..., J$; such that $J = 33$, representing the strategic variables to be optimized within feasible constraints – which can be implemented as $+/-1.0$ of the current values. Accordingly, the individual-level improvement at any x^* over the current feasible policy $(\mathbf{x}^{(str)})$ can be expressed as:

$$I_i(\mathbf{x}^{*(str)}, \mathbf{x}^{(str)}) = \max\left\{0, \frac{z^*_{ikt} - z_{ikt}}{K(\mathbf{x}^{*(str)}, \mathbf{x}^{(str)})^{-1}}\right\} \mathbb{I}\left\{c(\mathbf{x}^{*(str)}) > 0\right\} \tag{3.16}$$

The proposed acquisition function differs from extant improvement functions used in Bayesian optimization through the division of the inverse kernel evaluation, $K(x^*, x)^{-1}$. Intuitively, the denominator term penalizes any policies in the strategic variables that deviate too far from the current policy. Instead, the improvement function rewards policies that give *the paths of least resistance* to be executed by the marketing organization. Taken together, the final EI acquisition function is given by the following Monte Carlo integration (over both customers and the posterior distribution in brand affinity):

$$\alpha_{EI}(\mathbf{x}^{(str)}) = \sum_{i=1}^{N} E_{p(z)}[I_i(\mathbf{x}^{*(str)}, \mathbf{x}^{(str)})] \tag{3.17}$$

Through this approach, the Digital Marketing Twins framework can serve as a policy simulator, providing a mechanism to draw from the posterior predictive distribution of a customer profile across time and under different firms, in terms of their target satisfaction variables. This approach offers a proactive and strategic tool for rationalizing and optimizing a carrier's marketing strategies in a realistic and relevant way, which ultimately should improve customer satisfaction.

3.7 Conclusion

The novel Digital Marketing Twins methodology promises to address the challenges of analyzing large-scale customer surveys, as demonstrated in the context of the U.S.

wireless telecommunications retail market. This study thus addresses two major issues: (1) the theoretical difficulty of integrating customer surveys into a prescriptive framework and (2) the practical problem posed by repeated cross-sectional surveys, such that itcontributes to the literature on digital twins, machine learning methods for competitive environments, and customer satisfaction.

The proposed methodology provides counterfactual responses under different scenarios, which can serve as a powerful tool in the realm of customer analytics. The technique also addresses the missing data problem that is typical of repeated cross-sectional surveys, thereby presenting a comprehensive approach to understanding and leveraging customer survey data at scale.

The implementation of the methodology involves the development of a deep generative and probabilistic latent factor model that captures customer-side brand affinity at the individual level, for each brand and each time period, while controlling for observed heterogeneity and firm-side factors. The methodology leverages Bayesian optimization to maximize individual-level, latent, customer-side brand affinity, thereby leading to a "path of least resistance" at the individual level.

The findings have implications for marketers who seek to improve customer satisfaction by understanding the causes of satisfaction from surveys. Furthermore, because the methodology appears generalizable to other sectors and contexts, and therefore, it suggests new avenues for research and applications in the field of marketing and customer analytics.

Appendix A

Privacy Preserving Data Fusion

Deviations Between Observed and Imputed Response Distributions

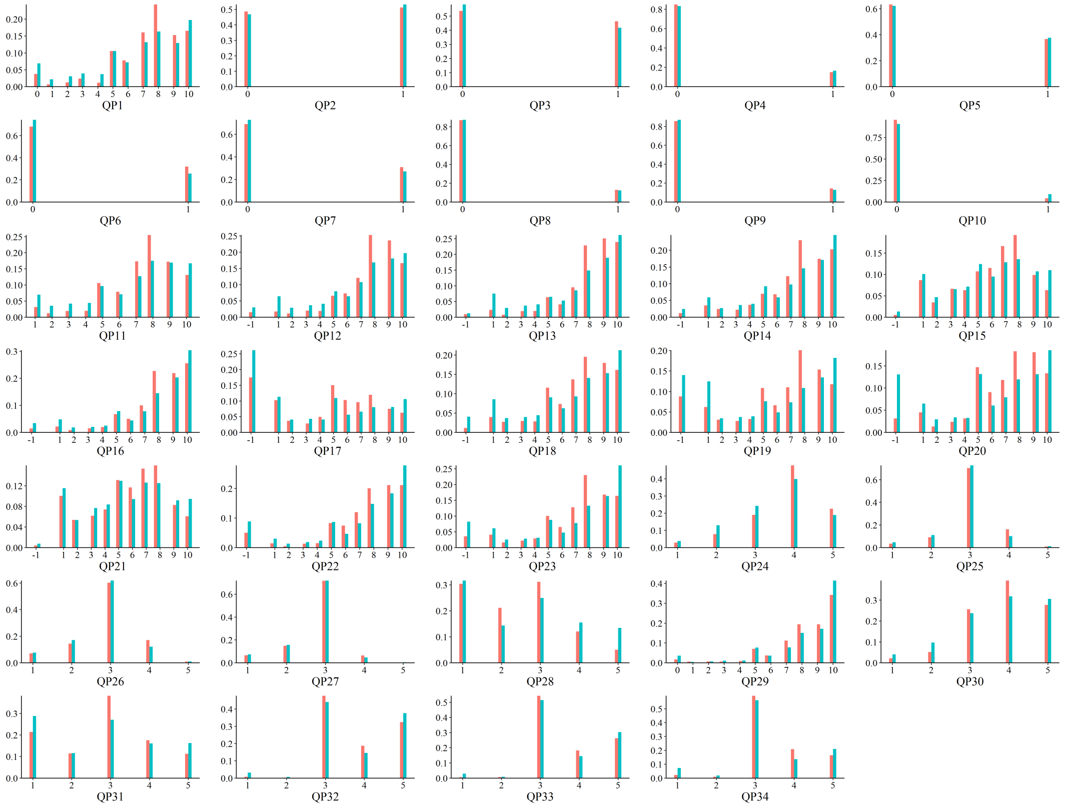


FIGURE A.1: Predicted responses of full customer base, in red, as imputed by PPDF, along with the observed responses of the 8K external survey respondents (in blue). The differences highlight the importance of correcting for selection bias before using insights from this and similar surveys. Survey respondents are more extreme in their Likelihood to Recommend.

A.1 Missing Data

Past techniques for handling missing data involved either removing observations if they had even one missing variable or filling missing values with the observed sample mean of the given variable(Graham, 2009). These techniques have been shown to be both inefficient and inaccurate. One drawback is that they reduce the sample size and could possibly create a non-random sample. They could also lead to inaccurate inferences due to the fact that using the mean of an explanatory variable can change the impact on the explained outcome.

Two more contemporary common methods are used to handle missing data in the social sciences in general, and in marketing research in particular. The first group is comprised of multiple imputation methods; key examples are described by Little and Rubin, 1989 and Kamakura and Wedel, 1997. The second group of methods for handling missing data are maximum likelihood methods. Such methods can be based on classical maximization of a model likelihood, as Kamakura and Wedel, 2000 demonstrate, or on stochastic simulation, such as Bayesian estimation used by Eleanor McDonnell Feit, Beltramo, and Feinberg, 2010.

Qian and Xie, 2014 proposed a Bayesian approach for completing missing data in regression covariates. A major contribution of their work is the ability to derive the missing values and regress over all data, without the need to specify the exact distribution for each covariate or the relations among covariates. This technique can handle high-dimensional missing covariate problems. However, when there is insufficient information to recover the underlying model, and when there is insufficient data or a too complex problem to handle, this method may not be suitable.

There are several types of missingness that should be acknowledged and properly handled. Specifically, Rubin, 1976 classifies three mechanisms of missing data: missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR, also known as non-ignorable).

In the case of MAR, while some of the data are missing, the missingness can be overcome by other observed variables. This means that the cause of missingness may depend on other covariates that are in the data but not on any unobserved data. A simple example can be responses to an income question on a survey. People might be reluctant to fill in their income if they feel that they have too high or too low of an income. However, missing answers can be imputed using a combination of other variables such as level of education, living area, and age. Though imputation will probably not arrive at the true individual responses, inferences drawn on large data samples will not be affected by the missingness.

MCAR can be considered a special case of MAR. As implied by the name, the missingness is completely random, and does not depend on either observed or unobserved variables. One possible example is when some of the data are corrupted due to a technical error. Another example is when some respondents simply forget to answer certain questions in a randomly-ordered questionnaire. Since they were randomly missing, any inferences regarding the data as a whole will be correct[1], thus representing the true underlying data generating processes, with or without completing the data.

The MNAR (also known as NR, nonrandom, or non-ignorable) missingness mechanism occurs when some of the values are missing and their missingness depends on unobserved data. Therefore, some information in the missing data depends on the missing values themselves, and cannot be fully rectified based on available information. One case is if inferences made based on such data might be biased by the missingness. Consider our example from above regarding missing income: if people are reluctant to respond to an income question because they are concerned about scammers, but such concern cannot be explained by any data available to us, then the missingness is non-random. It could be reverted to MAR if we could somehow account for this missing piece of information, perhaps through a survey question about this particular concern.

In our method, we will allow the data to have missingness of types MAR and MCAR. We will nonparametrically complete (augment) a latent representation of both within-variable missingness and the obvious whole variable missingness, with observed variables from either dataset. We define common variables as those that appear in both datasets and are measured on the same scale. Semi-common variables refer to variables that relate to the same underlying information, but are measured differently.

Given that missingness can occur across multiple data variables, a key limitation to conventional data imputation methods is that model complexity scales with the number of missing values. VAEs overcome this limitation by treating missingness as arising from a single generative model, and instead seek to encode the joint data generating process as a nonparametric random function that may in turn then be used to decode missing values when missingness may arise. The method allows for data – either individual values or entire covariates – to be missing at random (MAR), and for truncation into categories if data are semi-common.

Moreover, if data are MNAR, but the missingness can be accounted for (becoming MAR) using the other dataset, PPDF will be able to account for it as well. Consider a self-selected survey as our illustrated dataset. Customers who are more extreme in their attitudes towards the brand may be more likely to respond to a survey, and this attitudinal difference then translates into a distribution shift in terms of common variables. PPDF will be able to overcome such missingness, as long as the common variables bridge the missingness, essentially making MNAR on a single dataset become MAR in the joint dataset.

[1] With a caveat: the standard errors are likely to become larger, since we have fewer observations

A.2 Simulations: Sensitivity to Tuning Parameters

We begin with the sensitivity of the tuning parameters of the MNIST digits. As seen in Figure A.2, panels (b) and (c), while too small latent representation in Z may result in greater loss due to the inability to encode the data well enough, a value of Z which is too big may result in over-fitting, and might also result in higher reconstruction loss. Hidden layer dimensions may allow for richer representation, but they come at the cost of higher running times.

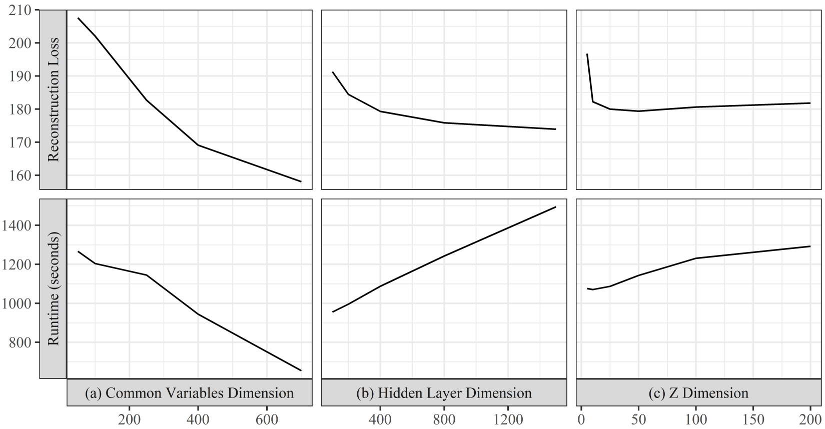


FIGURE A.2: Reconstruction loss (the loss relative to the original MNIST images, upper panels) and running time (lower panels) as a function of the number of common variables (panel (a)), length of the hidden layer (panel (b)) and length of Z – the latent representation of the encoders. The learning rate is $\eta = 2 \cdot 10^{-4}$. All simulations ran for 25 epochs. Number of common variables, if not varied, is 300 of the 784 pixels in each digit. Hidden layer dimension, if not varied, is 400. Latent representation – Z dimension – if not varied, is 25. An interesting artifact of VAEs, which is seen in the upper right panel, is that an increase in the size of the Z dimension may cause over-parameterization, which makes the reconstruction loss plateau or shift upwards.

Moving to the next sensitivity analysis, of the survey data described in Section 1.4.2, Figure A.3 shows the mean absolute error relative to the original data, across all columns of the joint dataset, after running for 10K epochs.

- Number of normalizing flows: Moving from left to right in the panels of Figure A.3, we can see the effects of the normalizing flow. The more flexibility there is in encoding the data (manifested through the number of flows), the better (lower) the mean absolute error. However, at some point, there is not much further improvement, and therefore, in the following stages, we utilize normalizing flows with 6 transform layers (a 9-flow iteration was also tested but found to have a similar plateau relative to the 6-flow iteration, so it was removed for brevity).

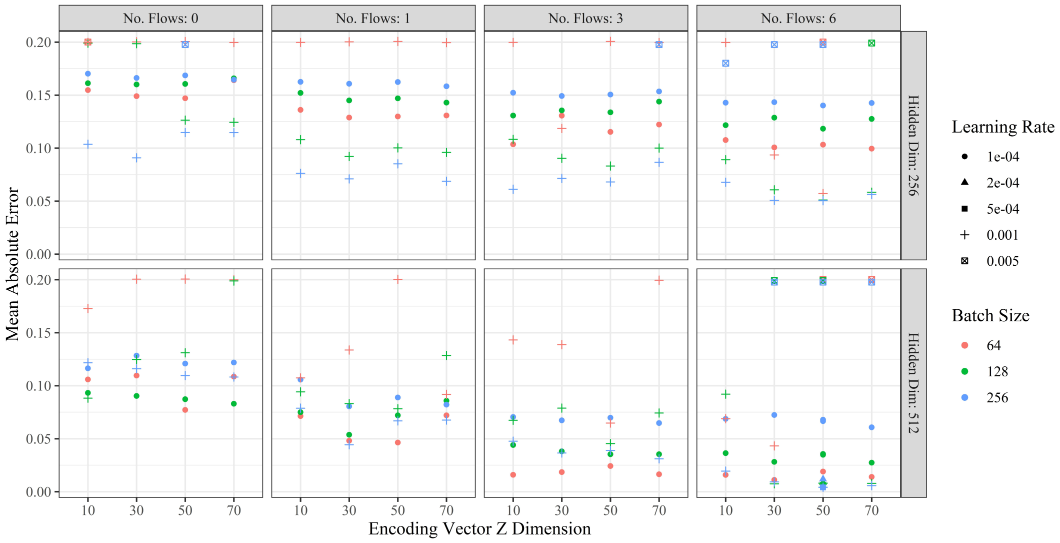


FIGURE A.3: Simulation results: Mean absolute error of data fusion without differential privacy, as a function of the dimension of the encoding vector Z (x axis), number of normalizing flows (horizontal panes), hidden layer dimension (vertical panes), learning rate (shape) and batch sizes (colors). A bigger hidden dimension is better, a bigger batch size is better, and a smaller learning rate is better in improving mean absolute error. Higher levels of the encoding vector Z dimension are usually better but may results in overfitting. More normalizing flows were also found to be better, but this value plateaued as well. Missing dots are due to simulations with these specifications failing to complete 10K epochs.

- Latent Encoding Vector Z: The length of the encoding vector has an inverse-U shape, where a smaller length is not sufficient to properly code the data into latent representation, but bigger sizes may result in overfitting the training data (this is seen visually by the loss that plateaus or shifts upward). Therefore, going forward, we use Z of size 50.

- Learning Rate: While largely beyond the scope of this paper, a smaller learning rate increases the likelihood of completing the training due to the gradual improvement rate but will result in a much longer time to train the model. We continued with a learning rate of 1e-4 for the sensitivity analyses of PPDF on this dataset.

- Batch Size: Each epoch divides the data into equal batches of pre-specified size. We find that the bigger the batch, the more accurate the analysis (up to a limit) and the faster the run. We carry on with a batch size of 256 rows. The batch size has a major effect on the run-time, though all runs in this analysis were efficient and ended in several minutes on a commodity GPU Google Colab Framework.

A.3 Survey Questions and CRM Data

The following represents a subset of the questions and variables in the data, presented in a way that preserves intellectual property of the respective companies. For further information, please contact the authors.

1. **Identifiers / Relationship Questions**

   - QI1: Do you identify as Hispanic or Latin?
   - QI2: Which wireless service providers do you currently use? (Select all that apply)
   - QI3: What is the current monthly cost for your primary personal phone with *[Respondent's Current Provider]*? If you have multiple phones, please specify the cost for the phone you use most frequently.
   - QI4: What was the main reason that led you to choose *[Respondent's Current Provider]* as your wireless service provider among the following options?
   - QI5: Which of the following products or services, if any, do you currently use or subscribe to in addition to your wireless service with *[Respondent's Current Provider]*?
   - QI6: What is the highest level of education you have completed?
   - QI7: What is the monthly data limit of your plan with *[Respondent's Current Provider]* before incurring additional charges for over usage?
   - QI8: Do you have access to *[Respondent's Current Provider]* 5G network service in the area where you live, work or frequently travel to, to the best of your knowledge?
   - QI9: What is the brand of the mobile phone that you currently use for personal use?
   - QI10: Which internet and cable/video providers do you primarily use for each service at home, among the following options?

2. **Engagement Questions**

   - QE1: What is the probability of you switching wireless service providers within the next 12 months?
   - QE2: In the context of considering switching carriers, which activities have you personally experienced with *[Respondent's Current Provider]* in the past 6 months?
   - QE3: In the context of asking about or disputing a recurring bill, which activities have you personally experienced with *[Respondent's Current Provider]* in the past 6 months?

- QE4: In the context of buying a new phone or connected device, which activities have you personally experienced with *[Respondent's Current Provider]* in the past 6 months?
- QE5: In the context of getting technical support (other than fixing a broken or lost device), which activities have you personally experienced with *[Respondent's Current Provider]* in the past 6 months?
- QE6: In the context of redeeming rewards, which activities have you personally experienced with [Telecom Carrier Name] in the past 6 months?
- QE7: In the context of fixing or replacing a broken, lost, or stolen device, which activities have you personally experienced with [Telecom Carrier Name] in the past 6 months?
- QE8: Other than the options provided, which activities have you personally experienced with [Telecom Carrier Name] in the past 6 months?
- QE9: During the past 6 months, have you used [Telecom Carrier Name]'s international plan?

3. **Perception Questions**

- QP1: How likely would you be to recommend your current wireless service provider, *[Respondent's Current Provider]*, to a friend or family member on a scale of 0-10?
- QP2: What network-related factors contributed to you giving your current provider, *[Respondent's Current Provider]*, a rating of *[Numerical Response to QP1*?
- QP3: What price-and-value-related factors contributed to you giving *[Respondent's Current Provider]* a rating of *[Numerical Response to QP1*?
- QP4: What billing-process-related factors contributed to you giving *[Respondent's Current Provider]* a rating of *[Numerical Response to QP1*?
- QP5: What customer-service-related factors contributed to you giving *[Respondent's Current Provider]* a rating of *[Numerical Response to QP1*?
- QP6: What general-feeling-related factors contributed to you giving *[Respondent's Current Provider]* a rating of *[Numerical Response to QP1*?
- QP7: What plan-related factors contributed to you giving *[Respondent's Current Provider]* a rating of *[Numerical Response to QP1*?
- QP8: What reward-and-benefit-related factors contributed to you giving *[Respondent's Current Provider]* a rating of *[Numerical Response to QP1*?
- QP9: What device-related factors contributed to you giving *[Respondent's Current Provider]* a rating of *[Numerical Response to QP1*?
- QP10: What other factors contributed to you giving *[Respondent's Current Provider]* a rating of *[Numerical Response to QP1*?

- QP11: What is your overall level of satisfaction with *[Respondent's Current Provider]*?
- QP12: How satisfied are you with the network speed provided by *[Respondent's Current Provider]* as part of your overall wireless service experience?
- QP13: How satisfied are you with the network reliability provided by *[Respondent's Current Provider]* as part of your overall wireless service experience?
- QP14: How satisfied are you with *[Respondent's Current Provider]*'s ability to provide data plans that meet your needs as part of your overall wireless service experience?
- QP15: How satisfied are you with the value you are receiving for the price you are paying for *[Respondent's Current Provider]*'s wireless service?
- QP16: How satisfied are you with the accuracy of billing provided by *[Respondent's Current Provider]* as part of your overall wireless service experience?
- QP17: How satisfied are you with the rewards and recognition offered by *[Respondent's Current Provider]* as part of your overall wireless service experience?
- QP18: How satisfied are you with the ease of doing business with *[Respondent's Current Provider]* as part of your overall wireless service experience?
- QP19: How satisfied are you with *[Respondent's Current Provider]*'s ability to solve problems on the first contact as part of your overall wireless service experience?
- QP20: How satisfied are you with the perception that *[Respondent's Current Provider]* is a brand for you as part of your overall wireless service experience?
- QP21: How satisfied are you with the total cost of your wireless service provided by *[Respondent's Current Provider]* as part of your overall wireless service experience?
- QP22: How satisfied are you with the device selection offered by *[Respondent's Current Provider]* as part of your overall wireless service experience?
- QP23: How satisfied are you with the way *[Respondent's Current Provider]* treats you fairly and respectfully as part of your overall wireless service experience?
- QP24: What is your overall feeling about [Telecom Carrier Name] as a wireless service provider?
- QP25: What is your overall feeling about [COMPETITOR A] as a wireless service provider?

- QP26: What is your overall feeling about [COMPETITOR B] as a wireless service provider?
- QP27: What is your overall feeling about Sprint as a wireless service provider?
- QP28: What is the likelihood of you making any changes (plan, provider, device) to your current wireless service in the next 6 months?
- QP29: What is the likelihood of you recommending your [pipe:QI9 (*sic*: read "Insert the phone brand that was given at question QI9")] phone to a friend or colleague?
- QP30: How well do you understand the details of your wireless plan with *[Respondent's Current Provider]*, on a scale of 1 to 5?
- QP31: How, if at all, has the coronavirus pandemic affected your spending habits?
- QP32: How has the coronavirus pandemic affected your perspective on the importance of home internet services?
- QP33: How has the coronavirus pandemic affected your perspective on the importance of wireless services?
- QP34: How has the coronavirus pandemic affected your perspective on the importance of TV and streaming services?

4. **Common variables**

- QC1: Can you please indicate your gender?
- QC2: Could you please provide your annual household income before taxes?
- QC3: Can you tell me what the total bill for your wireless plan was last month, exclusive of any charges for your device?
- QC4: Are you currently utilizing 5G service on a 5G compatible mobile phone?
- QC5: In the last 6 months, have you been to a *[Telecom Carrier Name]* retail location?
- QC6: Which provider do you primarily use for internet and cable/video service at home?
- QC7: In addition to your wireless service with *[Telecom Carrier Name]*, which, if any, of the following products/services do you use or subscribe to?
- QC8: Does your plan with *[Telecom Carrier Name]* include unlimited data?
- QC9: When will you fully own your phone and no longer have to make payments to *[Telecom Carrier Name]* for your phone?

- QC10: Are you currently a member of any of the following groups?

5. **Variables only available in the CRM Data**

   - QUC1: Risk score of the line with the highest likelihood of churn.
   - QUC2: Median total account amount, excluding device payments, from the previous six months' bills.
   - QUC3: The deviation of the total billed amount for this month from the average bill over the past 3, 6, and 12 months.
   - QUC4: A binary indicator of whether the account had at least one Sales and Service Transaction in the Customer Service Channel (1 = yes, 0 = no).
   - QUC5: A binary indicator of whether the account has an autopay discount (1 = yes, 0 = no).
   - QUC6: The count of 3G or 4G phones associated with the account.
   - QUC7: The most recent Quality Experience Score within the account that was lowest within the 90 days preceding the survey.
   - QUC8: Whether the feature was added within 180 days of the survey.
   - QUC9: Whether the account was disconnected (1 = yes, 0 = no).

Appendix B

Understanding Consumer Expenditure Through Gaussian Process Choice Models

B.1 Derivation of the Likelihood Function

After rearranging the terms, and denoting $g(.)$ as the standard extreme value density function and $G(.)$ as the standard extreme value cumulative distribution function:

$$L(q_1^* > 0, q_2^* > 0, \dots, q_M^* > 0, q_{M+1}^* = 0, \dots, q_J^* = 0) \tag{B.1}$$

$$= |\boldsymbol{J}| \int_{\varepsilon_1=-\infty}^{\varepsilon_1=+\infty} \left(\prod_{j=2}^{M} \frac{1}{\sigma} g\left(\frac{V_1 - V_j + \varepsilon_1}{\sigma} \right) \right) \left(\prod_{j=M+1}^{J} G\left(\frac{V_1 - V_j + \varepsilon_1}{\sigma} \right) \frac{1}{\sigma} g\left(\frac{\varepsilon_1}{\sigma} \right) \right) d\varepsilon_1$$

$$\begin{aligned} &= |\boldsymbol{J}| \int_{\varepsilon_1=-\infty}^{\varepsilon_1=+\infty} \left(\prod_{j=2}^{M} \frac{1}{\sigma} \exp\left(-\frac{V_1 - V_j + \varepsilon_1}{\sigma} - \exp\left(-\frac{V_1 - V_j + \varepsilon_1}{\sigma} \right) \right) \right) \\ &\quad \left(\prod_{j=M+1}^{J} \exp\left[-\exp\left[-\left(\frac{V_1 - V_j + \varepsilon_1}{\sigma} \right) \right] \right] \frac{1}{\sigma} \exp\left(-\frac{\varepsilon_1}{\sigma} - \exp\left(-\frac{\varepsilon_1}{\sigma} \right) \right) \right) d\varepsilon_1 \end{aligned} \tag{B.2}$$

$$\begin{aligned} &= |\boldsymbol{J}| \int_{\varepsilon_1=-\infty}^{\varepsilon_1=+\infty} \left(\prod_{j=2}^{M} \frac{1}{\sigma} \exp\left(-\frac{V_1 - V_j + \varepsilon_1}{\sigma} \right) * \left(-\exp\left(-\frac{V_1 - V_j + \varepsilon_1}{\sigma} \right) \right) \right) \\ &\quad \left(\prod_{j=M+1}^{J} \exp\left[-\exp\left[-\left(\frac{V_1 - V_j + \varepsilon_1}{\sigma} \right) \right] \right] \frac{1}{\sigma} \exp\left(-\frac{\varepsilon_1}{\sigma} \right) * \exp\left(-\exp\left(-\frac{\varepsilon_1}{\sigma} \right) \right) \right) d\varepsilon_1 \end{aligned} \tag{B.3}$$

$$\begin{aligned} &= |\boldsymbol{J}| \frac{1}{\sigma^{M-1}} \left[\prod_{j=2}^{M} \exp\left(-\left(\frac{V_1 - V_j}{\sigma} \right) \right) \right] \int_{\varepsilon_1=-\infty}^{\varepsilon_1=+\infty} \left(\exp\left(-\frac{\varepsilon_1}{\sigma} \right) \right)^{M-1} \\ &\quad \prod_{j=1}^{J} \exp\left(-\exp\left(-\left(\frac{V_1 - V_j + \varepsilon_1}{\sigma} \right) \right) \right) \frac{1}{\sigma} \exp\left(-\frac{\varepsilon_1}{\sigma} \right) d\varepsilon_1 \end{aligned} \tag{B.4}$$

$$\begin{aligned} &= |\boldsymbol{J}| \frac{1}{\sigma^{M-1}} \left[\prod_{j=2}^{M} \exp\left(-\left(\frac{V_1 - V_j}{\sigma} \right) \right) \right] \int_{\varepsilon_1=-\infty}^{\varepsilon_1=+\infty} \left(\exp\left(-\frac{\varepsilon_1}{\sigma} \right) \right)^{M-1} \\ &\quad \exp\left(-\sum_{j=1}^{J} \exp\left(-\left(\frac{V_1 - V_j + \varepsilon_1}{\sigma} \right) \right) \right) \frac{1}{\sigma} \exp\left(-\frac{\varepsilon_1}{\sigma} \right) d\varepsilon_1 \end{aligned} \tag{B.5}$$

Furthermore, we can write the following equality from the definition of a Gamma function:

$$
\begin{aligned}
&(M-1)! \\
&:= \Gamma(M) \\
&= \int_{z=0}^{z=+\infty} z^{M-1} \exp(-z) dz && \text{(B.6)} \\
&= \int_{t \sum_{j=1}^{J} \exp\left(-\frac{V_1-V_j}{\sigma}\right)=0}^{t \sum_{j=1}^{J} \exp\left(-\frac{V_1-V_j}{\sigma}\right)=+\infty} \left(t \sum_{j=1}^{J} \exp\left(-\frac{V_1-V_j}{\sigma}\right)\right)^{M-1} \exp\left(-t \sum_{j=1}^{J} \exp\left(-\frac{V_1-V_j}{\sigma}\right)\right) \\
&d\left(t \sum_{j=1}^{J} \exp\left(-\frac{V_1-V_j}{\sigma}\right)\right) && \text{(B.7)} \\
&= \left(\sum_{j=1}^{J} \exp\left(-\frac{V_1-V_j}{\sigma}\right)\right)^{M} \int_{t=0}^{t=+\infty} t^{M-1} \left(\sum_{j=1}^{J} \exp\left(-\frac{V_1-V_j}{\sigma}\right)\right)^{-1} \\
&\exp\left(-t \sum_{j=1}^{J} \exp\left(-\frac{V_1-V_j}{\sigma}\right)\right) \left(\sum_{j=1}^{J} \exp\left(-\frac{V_1-V_j}{\sigma}\right)\right) dt && \text{(B.8)} \\
&= \left(\sum_{j=1}^{J} \exp\left(-\frac{V_1-V_j}{\sigma}\right)\right)^{M} \int_{t=0}^{t=+\infty} t^{M-1} \exp\left(-t \sum_{j=1}^{J} \exp\left(-\frac{V_1-V_j}{\sigma}\right)\right) dt && \text{(B.9)}
\end{aligned}
$$

which implies:

$$
\frac{(M-1)!}{\left(\sum_{j=1}^{J} \exp\left(-\frac{V_1-V_j}{\sigma}\right)\right)^{M}} = -\int_{t=+\infty}^{t=0} t^{M-1} \exp\left(-t \sum_{j=1}^{J} \exp\left(-\frac{V_1-V_j}{\sigma}\right)\right) dt \qquad \text{(B.10)}
$$

A change of variable $t = \exp\left(-\frac{\varepsilon_1}{\sigma}\right)$ implies $dt = -\exp\left(-\frac{\varepsilon_1}{\sigma}\right)\frac{1}{\sigma} d\varepsilon_1$, which in turns implies, from the integration in (B.5):

$$
\begin{aligned}
&L(q_1^* > 0, q_2^* > 0, \ldots, q_M^* > 0, q_{M+1}^* = 0, \ldots, q_J^* = 0) \\
&= \frac{1}{\sigma^{M-1}} |\boldsymbol{J}| \left(\frac{\prod_{j=2}^{M} \exp\left(-\left(\frac{V_1-V_j}{\sigma}\right)\right)}{\left(\sum_{j=1}^{J} \exp\left(-\frac{V_1-V_j}{\sigma}\right)\right)^{M}} \right) (M-1)! \\
&= \frac{1}{\sigma^{M-1}} |\boldsymbol{J}| \left(\frac{\prod_{j=1}^{M} \exp\left(\frac{V_j}{\sigma}\right)}{\left(\sum_{j=1}^{J} \exp\left(\frac{V_j}{\sigma}\right)\right)^{M}} \right) (M-1)! && \text{(B.11)}
\end{aligned}
$$

where $V_j = \log\left(\frac{\partial v_j}{\partial q_j}\right) - \log(p_j)$ for all $j = 1, ..., J$. We calculate the Jacobian matrix as follows:

$$\boldsymbol{J}_{kl} = \frac{\partial\left(V_1 - V_{k+1} + \varepsilon_1\right)}{\partial q_{l+1}} \qquad \text{for } k, l = 1, 2, \ldots, M-1 \tag{B.12}$$

$$= \frac{\partial\left(V_1 - V_{k+1}\right)}{\partial q_{l+1}} \qquad \text{for } k, l = 1, 2, \ldots, M-1 \tag{B.13}$$

$$= \frac{\partial V_1}{\partial q_{l+1}} - \frac{\partial V_{k+1}}{\partial q_{l+1}} \qquad \text{for } k, l = 1, 2, \ldots, M-1 \tag{B.14}$$

$$= \frac{\partial \log\left(\frac{\partial v_1\left(\frac{1}{p_1}(x - \sum_{j\neq 1} p_j q_j)\right)}{\partial q_1}\right)}{\partial q_{l+1}} - \frac{\partial \log\left(\frac{\partial v_{k+1}(q_{k+1})}{\partial q_{k+1}}\right)}{\partial q_{l+1}} \qquad \text{for } k, l = 1, 2, \ldots, M-1 \tag{B.15}$$

If $k = l$, for $k = 1, \ldots, M-1$,

$$\boldsymbol{J}_{kl} = -\frac{p_{k+1}}{p_1}\frac{\frac{\partial^2 v_1}{\partial q_1^2}}{\frac{\partial v_1}{\partial q_1}} - \frac{\frac{\partial^2 v_{k+1}}{\partial q_{k+1}^2}}{\frac{\partial v_{k+1}}{\partial q_{k+1}}} \tag{B.16}$$

If $k \neq l$, for $k = 1, \ldots, M-1$ and $l = 1, \ldots, M-1$

$$\boldsymbol{J}_{kl} = -\frac{p_{l+1}}{p_1}\frac{\frac{\partial^2 v_1}{\partial q_1^2}}{\frac{\partial v_1}{\partial q_1}} \tag{B.17}$$

Therefore we have the following Jacobian matrix:

$$\boldsymbol{J} = \begin{bmatrix} -\frac{p_2}{p_1}\frac{\frac{\partial^2 v_1}{\partial q_1^2}}{\frac{\partial v_1}{\partial q_1}} - \frac{\frac{\partial^2 v_2}{\partial q_2^2}}{\frac{\partial v_2}{\partial q_2}} & -\frac{p_3}{p_1}\frac{\frac{\partial^2 v_1}{\partial q_1^2}}{\frac{\partial v_1}{\partial q_1}} & \cdots & -\frac{p_M}{p_1}\frac{\frac{\partial^2 v_1}{\partial q_1^2}}{\frac{\partial v_1}{\partial q_1}} \\ -\frac{p_2}{p_1}\frac{\frac{\partial^2 v_1}{\partial q_1^2}}{\frac{\partial v_1}{\partial q_1}} & -\frac{p_3}{p_1}\frac{\frac{\partial^2 v_1}{\partial q_1^2}}{\frac{\partial v_1}{\partial q_1}} - \frac{\frac{\partial^2 v_3}{\partial q_3^2}}{\frac{\partial v_3}{\partial q_3}} & \cdots & -\frac{p_M}{p_1}\frac{\frac{\partial^2 v_1}{\partial q_1^2}}{\frac{\partial v_1}{\partial q_1}} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{p_2}{p_1}\frac{\frac{\partial^2 v_1}{\partial q_1^2}}{\frac{\partial v_1}{\partial q_1}} & -\frac{p_3}{p_1}\frac{\frac{\partial^2 v_1}{\partial q_1^2}}{\frac{\partial v_1}{\partial q_1}} & \cdots & -\frac{p_M}{p_1}\frac{\frac{\partial^2 v_1}{\partial q_1^2}}{\frac{\partial v_1}{\partial q_1}} - \frac{\frac{\partial^2 v_M}{\partial q_M^2}}{\frac{\partial v_M}{\partial q_M}} \end{bmatrix} \tag{B.18}$$

Using the matrix determinant lemma, we write the absolute determinant of the Jacobian $|\boldsymbol{J}|$ as follows:

$$|\boldsymbol{J}| = \left(\prod_{j=1}^{M} -\frac{\partial^2 v_j/\partial q_j^2}{\partial v_j/\partial q_j}\right)\left(\frac{1}{p_1}\sum_{j=1}^{M}\frac{p_j}{-\frac{\partial^2 v_j/\partial q_j^2}{\partial v_j/\partial q_j}}\right) \tag{B.19}$$

If $M = 1$, there is no satiation effects and the Jacobian terms drops out and the model collapses to the standard MNL model (C. R. Bhat, 2008). Finally, we can write the closed form expression of the likelihood, combining (B.11) and (B.19):

$$L(q_1^* > 0, q_2^* > 0, \ldots, q_M^* > 0, q_{M+1}^* = 0, \ldots, q_J^* = 0) = \\ \frac{1}{\sigma^{M-1}} \left(\prod_{j=1}^{M} - \frac{\partial^2 v_j / \partial q_j^2}{\partial v_j / \partial q_j} \right) \left(\frac{1}{p_1} \sum_{j=1}^{M} \frac{p_j}{- \frac{\partial^2 v_j / \partial q_j^2}{\partial v_j / \partial q_j}} \right) \left(\frac{\prod_{j=1}^{M} \exp\left(\frac{V_j}{\sigma}\right)}{\left(\sum_{j=1}^{J} \exp\left(\frac{V_j}{\sigma}\right)\right)^M} \right) (M-1)! \tag{B.20}$$

where $V_j = \log\left(\frac{\partial v_j(q_j)}{\partial q_j}\right) - \log(p_j)$ for all $j = 1, ..., J$, which is also equal to:

$$L(q_1^*, q_2^*, \ldots, q_M^*, 0, \ldots, 0) = \frac{1}{\sigma^{M-1}} \left(\prod_{j=1}^{M} \left| - \frac{\partial}{\partial q_j} \log\left(\frac{\partial v_j}{\partial q_j}\right) \right| \right) \left(\frac{1}{p_1} \left| \sum_{j=1}^{M} \frac{p_j}{- \frac{\partial}{\partial q_j} \log\left(\frac{\partial v_j}{\partial q_j}\right)} \right| \right) \tag{B.21}$$

$$\left(\frac{\prod_{j=1}^{M} \exp\left(\frac{V_j}{\sigma}\right)}{\left(\sum_{j=1}^{J} \exp\left(\frac{V_j}{\sigma}\right)\right)^M} \right) (M-1)!$$

where $V_j = \log\left(\frac{\partial v_j}{\partial q_j}\right) - \log(p_j)$ for all $j = 1, ..., J$.

B.2 Bayesian Nonparametric Prior Model for Unknown Monotonic Functions

We have seen that the Kuhn-Tucker first order conditions are necessary and sufficient when the utility function is strictly quasi-concave. When marginal utility functions are positive and increasing, then the utility function will be monotonic and strictly quasi-concave. However, Gaussian processes priors in their basic formulation are not restricted in their shape, i.e. they are not a priori positive and increasing. Thus, in order to build valid priors compatible with our microeconomic domain of expertise, we need to transform these priors and constraint their shape adequately. There are several ways to do so.

First, we could impose a set of virtual locations of the sign of the derivative of the process (Riihimäki and Vehtari, 2010). These points would form a grid of constraints. These points could enforce the monotonicity of the sub-utility levels through the positivity of the marginal sub-utilities evaluated at these points. However there are theoretical and empirical issues with this approach. Monotonicity of the sample paths may not be guaranteed on the entire domain. One question is to determine an appropriate number of virtual observations: too many points imply that the posterior will be overly smoothed and too few points imply non-monotonicity. Furthermore, the posterior and the marginal likelihood would then depend on virtual locations. Finally, optimal quantities in a given test set (or at predictive inference time) could be far away from optimal quantities in the training set (or inference time). As a consequence, virtual locations would need to be correctly placed, accounting for both existing optimal solutions and future solutions, which is not practical. For all these reasons, we stay away from the virtual location approach and propose an alternative solution to build monotonicity directly into our priors, developing an approach proposed by Andersen et al., 2018 and Solin and Särkkä, 2020. The approach is also empirically investigated in Riutort-Mayol et al., 2023.

Our approach to build a valid prior in the monotonic function space relies on a non-linear transformation of our original Gaussian process prior. Since our domain expertise informs us that the marginal utility for each good $j = 1, \ldots, J$ should be decreasing, equivalently, the logarithm of the marginal utility should also be decreasing. This monotonicity guarantees that the utility for each good j is concave and increasing, and implies that the utility function is strictly quasi-concave.

We characterize the set of non-increasing functions as the set of solutions to the differential equation

$$\frac{\partial}{\partial \boldsymbol{q}_j} \log \left(\frac{\partial \boldsymbol{\nu}_j}{\partial \boldsymbol{q}_j} \right) = -h \leq 0 \tag{B.22}$$

where $h(.)$ is a non-negative function. The solutions to this differential equation will be monotonic and given by

$$\log\left(\frac{\partial \boldsymbol{\nu}_j(q_j)}{\partial \boldsymbol{q}_j}\right) = f_0 - \int_a^x h(s)ds \tag{B.23}$$

for f_0 and a constants and $x \geq a$. We assume that h can be modeled by a nonlinear transformation applied to a Gaussian process:

$$h(x) = (t \circ g)(x) \tag{B.24}$$

where t is a non negative function and g is a sample from a GP. Conditioned on g, the log marginal utility function is given by

$$\log\left(\frac{\partial \boldsymbol{\nu}_j(q_j)}{\partial \boldsymbol{q}_j}\right) = f_0 - \int_a^x t(g(s))ds \tag{B.25}$$

Then by construction the log marginal utility will be absolutely continuous, and has the property that $\log\left(\frac{\partial \boldsymbol{\nu}_j(a)}{\partial \boldsymbol{q}_j}\right) = f_0$ and $\log\left(\frac{\partial \boldsymbol{\nu}_j(q_j)}{\partial \boldsymbol{q}_j}\right) \leq f_0$ for all values $x \geq a$. Therefore $\frac{\partial \boldsymbol{\nu}_j(q_j)}{\partial \boldsymbol{q}_j}$ will be bounded by $\exp(f_0)$. The function $t(.)$ can be any non-negative transformation, but in this work, we choose $t(x) = x^2$ for modeling convenience and tractability.

Next, we propose to apply a low rank approximation based on Hilbert space methods, to make inference tractable. The evaluation of the log marginal utility function is not trivial because it depends on all values of $g(s)$ for $a < s < x$ (Andersen et al., 2018). We want to evaluate a closed form approximation of the log marginal utility function. To arrive at this close form approximation, we assume that the domain of interest is a compact subset $x \in [-M, M]$ for some $M > 0$ and that g satisfies the Dirichlet boundary condition $g(-M) = g(M) = 0$. Then the stationary covariance function K for g can be approximated (Solin and Särkkä, 2020) as

$$K(q_j, q_j) \approx \sum_l S_\theta(\sqrt{\lambda})\phi_l(q_j)\phi_l(q_j') \tag{B.26}$$

where $S_\theta(.)$ is the spectral density of the stationary covariance function K and θ is the set of hyperparameters of K. Since we are using a squared exponential covariance function, we have the following closed form spectral density:

$$S_\theta(\omega) = \sigma_f^2\sqrt{2\pi}\beta\exp\left(-\frac{1}{2}\beta^2\omega^T\omega\right) \tag{B.27}$$

The eigenvalues $\{\lambda_l\}_{l=1}^{\infty}$ and eigenfunctions $\{\phi_l\}_{l=1}^{\infty}$ of the Laplacian operators satisfy the following eigenvalue problem

$$\begin{cases} -\nabla^2 \phi_l(q_j) & = \lambda_l \phi_l(q_j) \quad q_j \in (-M, M) \\ \phi_l(q_j) & = 0 \qquad\qquad q_j \notin (-M, M) \end{cases} \tag{B.28}$$

The eigenvalues $\lambda_j > 0$ are real and positive because the Laplacian operator is a positive definite Hermitian operator. If we truncate the sum in equation (B.26), we can represent g as follows

$$g(q_j) \approx \sum_{l=1}^{L} \alpha_l \phi_l(q_j) \tag{B.29}$$

where $\alpha_j \sim \mathcal{N}\left(0, S_\theta\left(\lambda_j^{1/2}\right)\right)$. In other words, we use an approximation of the function g with a finite basis function expansion, where each coefficient α_l is Gaussian distributed with zero mean and variance $S_\theta\left(\lambda_j^{1/2}\right)$. As an aside, an extensive empirical analysis on the choice of M and L is available in Riutort-Mayol et al., 2023. The eigenfunctions ϕ_l for the eigenvalue problem in equation (B.28):

$$\sqrt{\lambda_l} = \frac{l\pi}{2M} \tag{B.30}$$

$$\phi_l(q_j) = \sqrt{\frac{1}{M}} \sin\left(\sqrt{\lambda_l}(q_j + M)\right) \tag{B.31}$$

Then, we can substitute the approximation of $g(.)$ in equation (B.29) and $t(x) = x^2$ into equation (B.25)

$$\log\left(\frac{\partial \boldsymbol{\nu}_j(q_j)}{\partial \boldsymbol{q}_j}\right) \approx f_0 - \int_a^x \left[\sum_{l=1}^{L} \alpha_l \phi_l(s)\right]^2 ds \tag{B.32}$$

$$\log\left(\frac{\partial \boldsymbol{\nu}_j(q_j)}{\partial \boldsymbol{q}_j}\right) \approx f_0 - \int_a^x \sum_{l=1}^{L} \sum_{l'=1}^{L} \alpha_l \alpha_{l'} \phi_l(s) \phi_{l'}(s) ds \tag{B.33}$$

$$\approx f_0 - \sum_{l=1}^{L} \sum_{l'=1}^{L} \alpha_l \alpha_{l'} \int_a^x \phi_l(s) \phi_{l'}(s) ds \tag{B.34}$$

And because the eigenfunctions are sinusoids, the integrands, $\phi_l(s)\phi_{l'}(s)$ can be rewritten using the trigonometric identity: $\sin(a)\sin(b) = \frac{1}{2}(\cos(a-b) - \cos(a+b))$, which lead to analytical solvability of the definite integrals:

$$\log\left(\frac{\partial \boldsymbol{\nu}_j(q_j)}{\partial \boldsymbol{q}_j}\right) \approx f_0 - \sum_{l=1}^{L} \sum_{l'=1}^{L} \alpha_l \alpha_{l'} \psi_{ll'}(q_j)$$

$$\log\left(\frac{\partial \boldsymbol{\nu}_j(q_j)}{\partial \boldsymbol{q}_j}\right) = f_0 - \boldsymbol{\alpha}^T \boldsymbol{\psi}(q_j) \boldsymbol{\alpha} \tag{B.35}$$

where $\psi(q_j)$ is given by

$$\psi_{ll'}(q_j) = \begin{cases} \frac{1}{2M}(q_j + M) - \frac{\sin\left(\gamma^{+}_{l'l'}(q_l+M)\right)}{2M\gamma^{+}_{l'l'}}, & l = l' \\ \frac{\sin\left(\gamma^{-}_{ll'}(q_l+M)\right)}{2M\gamma^{-}_{ll'}} - \frac{\sin\left(\gamma^{+}_{ll'}(q_l+M)\right)}{2M\gamma^{+}_{ll'}}, & l \neq l' \end{cases} \tag{B.36}$$

where $\gamma^{-}_{ll'} = \sqrt{\lambda_l} - \sqrt{\lambda_l}$, $\gamma^{+}_{ll'} = \sqrt{\lambda_l} + \sqrt{\lambda_l}$ and $a = -M$. Under reasonable conditions, the kernel approximation in equation (B.26) becomes exact when $L \to \infty$ and $M \to \infty$ (Solin and Särkkä, 2020). Equation (B.36) shows that the functions $\psi_{ll'}$ are independent of the kernel hyperparameters and can be precomputed. Finally, we get a closed form approximation for the derivative of the log marginal utility by taking the derivative of equations (B.35) and (B.36):

$$\frac{\partial}{\partial q_j} \log\left(\frac{\partial \boldsymbol{\nu}_j(q_j)}{\partial \boldsymbol{q}_j}\right) = -\boldsymbol{\alpha}^T \frac{\partial}{\partial q_j} \boldsymbol{\psi}(q_j)\boldsymbol{\alpha} \tag{B.37}$$

and

$$\frac{\partial}{\partial q_j}\psi_{ll'}(q_j) = \begin{cases} \frac{1}{2M}\left[1 - \cos\left(\gamma^{+}_{l'l'}(q_l + M)\right)\right], & l = l' \\ \frac{1}{2M}\left[\cos\left(\gamma^{-}_{ll'}(q_l + M)\right) - \cos\left(\gamma^{+}_{ll'}(q_l + M)\right)\right], & l \neq l' \end{cases} \tag{B.38}$$

In conclusion, we have derived a closed form approximation for the prior distribution over two types of functions of interest: the monotonically decreasing log marginal utility function for each good $j = 1, \ldots, J$ and the respective partial derivative of that function. These expressions above can then be used within a generic derivative-based MCMC algorithm to compute the posterior distribution.

B.3 Demand Prediction Algorithm

In this section, we provide an algorithm for demand prediction given a certain utility specification. The algorithm borrows from Pinjari and C. Bhat (2021) and implements a numerical bisection to estimate the Lagrange multiplier, which is further used to calculate optimal quantities.

First, we propose a procedure to obtain an inverse function approximation of the marginal utility, using a fine sequence of points. This procedure can be applied for any strictly decreasing function, with predefined domain and range. Let j be the index of the j^{th} good and suppose that the optimal quantity q_j^* for this good is strictly positive. Then, according to the Kuhn-Tucker conditions, $\frac{\partial u_j(q_j^*)}{\partial q_j} = \lambda p_j$. Since $\frac{\partial u_j}{\partial q_j}$ is strictly decreasing, it has an inverse. Let $\left(\frac{\partial u_j}{\partial q_j}\right)^{-1}$ be the inverse of $\frac{\partial u_j}{\partial q_j}$. Then we have the following: $q_j^* = \left(\frac{\partial u_j}{\partial q_j}\right)^{-1}(\lambda p_j)$.

1. Let $\bar{q}_j$ be a regular sequence of points from $\bar{q}_{min}$ to $\bar{q}_{max}$, with increment $\bar{q}_{inc}$. For example, one could set: $\bar{q}_{min} = 10^{-3}$ to $\bar{q}_{max} = 200$, with increment $\bar{q}_{inc} = 10^{-3}$. Let C_j be the cardinal number of $\bar{q}_j$. Let an element of $\bar{q}_j$ be q_{ji} for $i = 1, \ldots, C_j$.

2. Sort all values of $\bar{q}_{ji}$ in a descending order. Let's call this tuple $R(\bar{q}_j)$. Sort all values of $\frac{\partial u_j(\bar{q}_{ji})}{\partial q_j}$, for all $i = 1, \ldots, C_j$ in an ascending order. Let's call this tuple $S(\bar{q}_j)$. Let $R_i(\bar{q}_j)$ and $S_i(\bar{q}_j)$ respectively denote the i^{th} element in $R(\bar{q}_j)$ and $S(\bar{q}_j)$.

3. Let $\hat{\lambda}$ denote an estimate of λ, and p_j price of good j. Calculate $\hat{\lambda} p_j$ and search the index k of the first nearest value to $\hat{\lambda} p_j$ in $S(\bar{q}_j)$. Then we have the following approximation: $S_k(\bar{q}_j) \approx \hat{\lambda} p_j$.

4. Calculate $S_k(\bar{q}_j)$. Then we have the following approximation:
   $R_k(\bar{q}_j) \approx \left(\frac{\partial u_j}{\partial q_j}\right)^{-1}(\hat{\lambda} p_j) \approx q_j^*$.

Now, let $\hat{\lambda}$ and $\hat{x}$ be respectively estimates for the Lagrange multiplier λ and the budget constraint x. Let δ_λ and δ_x be the tolerance values fixed to small values such as 10^{-6}.

The algorithm for demand prediction goes as follows:

1. Set M = 1, where M is the number of consumed goods at the optimum. Compute the price normalized deterministic marginal utility values at zero consumption for all J alternatives, $\frac{\partial v_j(0)/\partial q_j}{p_j}$, for $j = 1, \ldots, J$. Simulate independent errors ε_j from an Extreme Value distribution with scale σ for each good, and multiply them respectively to the J alternatives: $\frac{\partial u_j(0)/\partial q_j}{p_j} = \frac{\partial v_j(0)/\partial q_j}{p_j} * \exp(\varepsilon_j)$. Sort the J stochastic values in a descending order. If there is an outside good, it is positioned first.

2. Set $\hat{\lambda} = \frac{\partial u_{M+1}(0)/\partial q_{M+1}}{p_{M+1}}$. For all $m = 1, \ldots, M$, using $\hat{\lambda}$ and p_m, get an estimate of the demand $\hat{q}_m$ using the inverse function approximation described above, following from a consequence of the Kuhn-Tucker conditions: $\hat{q}^*_m = \left(\frac{\partial u_m}{\partial q_m}\right)^{-1}(\hat{\lambda} p_m)$. Get an estimate of the budget constraint x, using
$\hat{x} = \sum_{m=1}^{M} p_j \hat{q}^*_m$.

3. If $\hat{x} < x$, go to step 4. Else, if $\hat{x} > x$, set $\lambda_L = \frac{\partial u_{M+1}(0)/\partial q_{M+1}}{p_{M+1}}$ and $\lambda_U = \frac{\partial u_M(0)/\partial q_M}{p_M}$ because $\frac{\partial u_{M+1}(0)/\partial q_{M+1}}{p_{M+1}} < \lambda < \frac{\partial u_M(0)/\partial q_M}{p_M}$. Go to step 5 to estimate λ with numerical bisection.

4. Set M = M + 1. If $M < J$, then go to step 2. Else, if $M = J$, set $\lambda_L = 0$ and $\lambda_U = \frac{\partial v_J(0)/\partial q_J}{p_J}$ since $0 < \lambda < \frac{\partial v_J(0)/\partial q_J}{p_J}$.

5. Set $\hat{\lambda} = \frac{\lambda_L + \lambda_U}{2}$. Re-calculate a new estimate of the demand $\hat{q}_m$ for all $m = 1, \ldots, M$ as in step 2, and get a new estimate of the budget constraint $\hat{x}$.

   (a) If $|\lambda_L - \lambda_U| \leq \delta_\lambda$ or $|\hat{x} - x| \leq \delta_x$ then go to step 6.

   (b) Else, if $\hat{x} < x$ update the upper bound of λ: $\lambda_U = \frac{\lambda_U + \lambda_L}{2}$ and go to step 5(a).

   (c) Else, if $\hat{x} > x$ update the lower bound of λ: $\lambda_L = \frac{\lambda_U + \lambda_L}{2}$ and go to step 5(a).

6. Compute the optimal quantities of the first M goods using the method in Step 2. Set the optimal quantities of the other goods to zero and stop.

Appendix C

Digital Twins: A Generative Approach for Counterfactual Customer Analytics

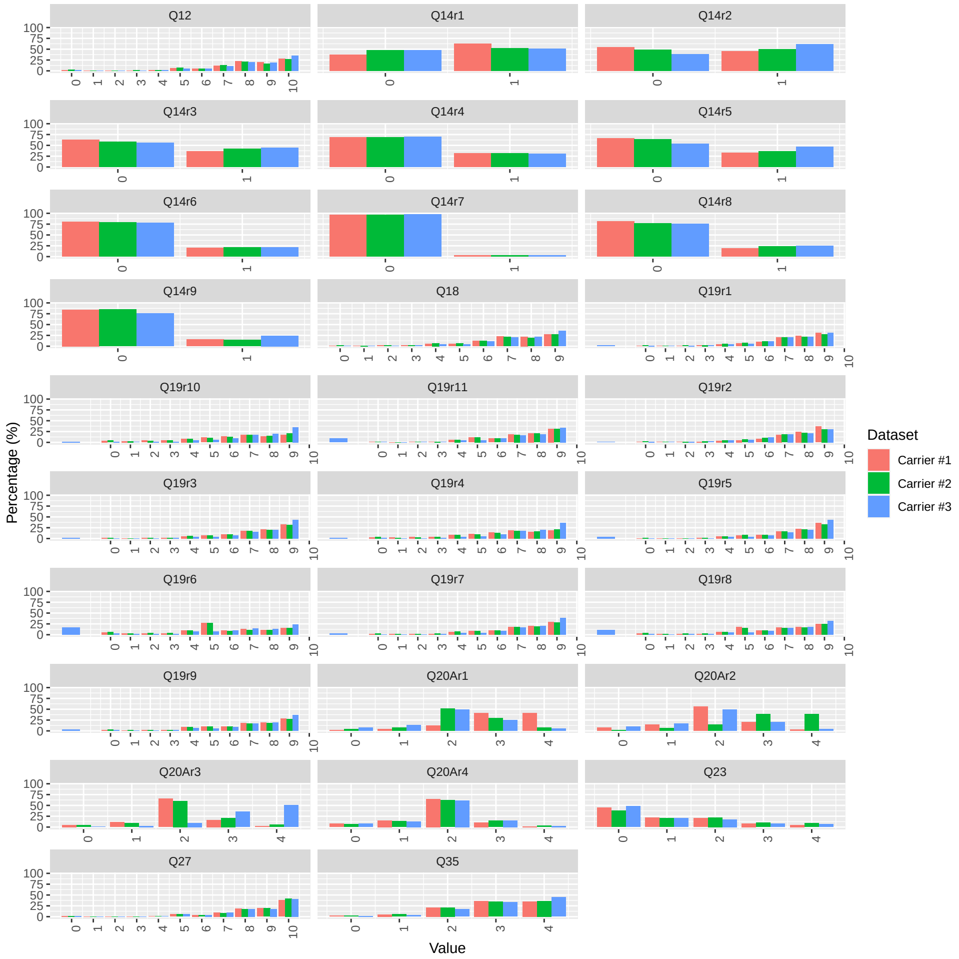


FIGURE C.1: Summary Statistics for Target Variables and Strategic Variables, Per Carrier

Variable Name	Description
Q14R1	[pipe:hCurrentProvider]'s Network - Which of the following directly contributed to you giving [pipe:hCurrentProvider] a rating of ?
Q14R2	[pipe:hCurrentProvider]'s Price / Value - Which of the following directly contributed to you giving [pipe:hCurrentProvider] a rating of ?
Q14R8	[pipe:hCurrentProvider]'s Billing process - Which of the following directly contributed to you giving [pipe:hCurrentProvider] a rating of ?
Q14R3	[pipe:hCurrentProvider]'s Customer Service - Which of the following directly contributed to you giving [pipe:hCurrentProvider] a rating of ?
Q14R4	General feelings about [pipe:hCurrentProvider] - Which of the following directly contributed to you giving [pipe:hCurrentProvider] a rating of ?
Q14R5	[pipe:hCurrentProvider]'s Plans - Which of the following directly contributed to you giving [pipe:hCurrentProvider] a rating of ?
Q14R9	[pipe:hCurrentProvider]'s Rewards and benefits - Which of the following directly contributed to you giving [pipe:hCurrentProvider] a rating of ?
Q14R6	[pipe:hCurrentProvider]'s Devices - Which of the following directly contributed to you giving [pipe:hCurrentProvider] a rating of ?
Q14R7	Other (please specify) - Which of the following directly contributed to you giving [pipe:hCurrentProvider] a rating of ?
Q19R1	Network speed - How satisfied are you with [pipe:hCurrentProvider]'s performance on the following aspects of your overall wireless service experience?
Q19R2	Network reliability - How satisfied are you with [pipe:hCurrentProvider]'s performance on the following aspects of your overall wireless service experience?
Q19R3	Data plans that meet my needs - How satisfied are you with [pipe:hCurrentProvider]'s performance on the following aspects of your overall wireless service experience?
Q19R4	Value for the price paid - How satisfied are you with [pipe:hCurrentProvider]'s performance on the following aspects of your overall wireless service experience?
Q19R5	Accuracy of billing - How satisfied are you with [pipe:hCurrentProvider]'s performance on the following aspects of your overall wireless service experience?
Q19R6	Rewards and recognition - How satisfied are you with [pipe:hCurrentProvider]'s performance on the following aspects of your overall wireless service experience?
Q19R7	Easy to do business with - How satisfied are you with [pipe:hCurrentProvider]'s performance on the following aspects of your overall wireless service experience?
Q19R8	Solves problems the first time you contact them - How satisfied are you with [pipe:hCurrentProvider]'s performance on the following aspects of your overall wireless service experience?
Q19R9	Is a brand for me - How satisfied are you with [pipe:hCurrentProvider]'s performance on the following aspects of your overall wireless service experience?
Q19R10	Total cost of wireless service - How satisfied are you with [pipe:hCurrentProvider]'s performance on the following aspects of your overall wireless service experience?
Q19R11	Device Selection - How satisfied are you with [pipe:hCurrentProvider]'s performance on the following aspects of your overall wireless service experience?
Q19R1aux (resp. Q19R2-Q19R11)	Aware of [pipe:hCurrentProvider]'s performance on the aspects mentioned in Q19R1? (resp. Q19R2-Q19R11)
Q23	How likely are you to change anything (plan, provider, device) about your current wireless service in the next 6 months?
Q35	On a scale of 1 to 5, how well do you feel you understand the details of your wireless plan with [pipe:hCurrentProvider]?

TABLE C.1: List of Strategic Variables

Variable Name	Description
Q12	Thinking about your overall experience with your wireless service provider, on a scale of 0 to 10, how likely are you to recommend [pipe:hCurrentProvider] to a friend or family member?
Q18	Q18: Overall, how satisfied are you with [pipe:hCurrentProvider]?
Q20	How likely are you to switch wireless service providers within the next 12 months?
Q20AR1	Carrier 1 - What best describes your overall feeling about each wireless service provider?
Q20AR2	Carrier 2 - What best describes your overall feeling about each wireless service provider?
Q20AR3	Carrier 3 - What best describes your overall feeling about each wireless service provider?
Q27	How likely are you to recommend your [pipe:Q26] phone to a friend or a colleague?

TABLE C.2: List of Target Variables

Bibliography

Abadi, Martin et al. (2016). “Deep learning with differential privacy”. In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318.

Allenby, Greg M, Mark J Garratt, and Peter E Rossi (2010). “A Model for Trade-Up and Change in Considered Brands”. In: *Marketing Science* 29.1, pp. 40–56.

Allenby, Greg M and Peter E Rossi (1991). “Quality Perceptions and Asymmetric Switching Between Brands”. In: *Marketing science* 10.3, pp. 185–204.

Anand, Piyush and Clarence Lee (2023). “Using Deep Learning to Overcome Privacy and Scalability Issues in Customer Data Transfer”. In: *Marketing Science* 42.1, pp. 189–207.

Andersen, MR et al. (2018). “A non-parametric probabilistic model for monotonic functions.“” In: *All Of Bayesian Nonparametrics” Workshop at NeurIPS*.

Ascarza, Eva (2018). “Retention futility: Targeting high-risk customers might be ineffective”. In: *Journal of Marketing Research* 55.1, pp. 80–98.

Ascarza, Eva et al. (2018). “In pursuit of enhanced customer retention management: Review, key issues, and future directions”. In: *Customer Needs and Solutions* 5.1, pp. 65–81.

Athey, Susan, Stefan Wager, et al. (2017). *Efficient policy learning*. Tech. rep.

Bagdasaryan, Eugene, Omid Poursaeed, and Vitaly Shmatikov (2019). “Differential privacy has disparate impact on model accuracy”. In: *Advances in neural information processing systems* 32.

Becker, Gary S (1965). “A Theory of the Allocation of Time”. In: *The economic journal* 75.299, pp. 493–517.

Berry, Steven, James Levinsohn, and Ariel Pakes (2004). “Differentiated products demand systems from a combination of micro and macro data: The new car market”. In: *Journal of political Economy* 112.1, pp. 68–105.

Betancourt, Michael (2020). *Ordinal Regression Case Study, section 2.2*. [Online; accessed 30-June-2023]. URL: https://betanalpha.github.io/assets/case_studies/ordinal_regression.html.

Bethlehem, Jelke (2010). “Selection bias in web surveys”. In: *International statistical review* 78.2, pp. 161–188.

Bhat, Chandra R (2005). “A multiple discrete–continuous extreme value model: formulation and application to discretionary time-use decisions”. In: *Transportation Research Part B: Methodological* 39.8, pp. 679–707.

Bhat, Chandra R (2008). "The Multiple Discrete-Continuous Extreme Value (MDCEV) Model: Role of Utility Function Parameters, Identification Considerations, and Model Extensions". In: *Transportation Research Part B: Methodological* 42.3, pp. 274–303.

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017). "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518, pp. 859–877.

Bottou, Léon et al. (2013). "Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising." In: *Journal of Machine Learning Research* 14.11.

Bradburn, Norman M et al. (1979). *Improving interview method and questionnaire design: Response effects to threatening questions in survey research*. Jossey-Bass San Francisco.

Bradlow, Eric T and Alan M Zaslavsky (1999). "A hierarchical latent variable model for ordinal data from a customer satisfaction survey with "no answer" responses". In: *Journal of the American Statistical Association* 94.445, pp. 43–52.

Braun, Michael and David A Schweidel (2011). "Modeling customer lifetimes with multiple causes of churn". In: *Marketing Science* 30.5, pp. 881–902.

Chandar, Sarath et al. (2016). "Correlational neural networks". In: *Neural computation* 28.2, pp. 257–285.

Chiang, Jeongwen (1991). "A Simultaneous Approach to the Whether, What and How Much to Buy Questions". In: *Marketing Science* 10.4, pp. 297–315.

Chintagunta, Pradeep K and Harikesh S Nair (2011). "Structural workshop paper — discrete-choice models of consumer demand in marketing". In: *Marketing Science* 30.6, pp. 977–996.

Clinton, Joshua, Simon Jackman, and Douglas Rivers (2004). "The statistical analysis of roll call data". In: *American Political Science Review* 98.2, pp. 355–370.

Dautov, Rustem, Salvatore Distefano, and Rajkumaar Buyya (2019). "Hierarchical data fusion for Smart Healthcare". In: *Journal of Big Data* 6.1, pp. 1–23.

De Haan, Evert, Peter C Verhoef, and Thorsten Wiesel (2015). "The predictive ability of different customer feedback metrics for retention". In: *International Journal of Research in Marketing* 32.2, pp. 195–206.

Deng, Li (2012). "The mnist database of handwritten digit images for machine learning research [best of the web]". In: *IEEE Signal Processing Magazine* 29.6, pp. 141–142.

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

Dew, Ryan, Asim Ansari, and Yang Li (2020). "Modeling Dynamic Heterogeneity Using Gaussian Processes". In: *Journal of Marketing Research* 57.1, pp. 55–77.

Dew, Ryan, Eva Ascarza, et al. (2023). "Detecting Routines: Applications to Ridesharing Customer Relationship Management". In: *Journal of Marketing Research*.

Dew, Ryan and Yuhao Fan (2021). "A Gaussian Process Model of Cross-Category Dynamics in Brand Choice". In: *arXiv preprint arXiv:2104.11702*.

Dias, Felipe F et al. (2019). "Fusing multiple sources of data to understand ride-hailing use". In: *Transportation Research Record* 2673.6, pp. 214–224.

Diggle, Peter J, Jonathan A Tawn, and Rana A Moyeed (1998). "Model-Based Geostatistics". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47.3, pp. 299–350.

Domingo-Ferrer, Josep and Vicenç Torra (2008). "A critique of k-anonymity and some of its enhancements". In: *2008 Third International Conference on Availability, Reliability and Security*. IEEE, pp. 990–993.

Dubé, Jean-Pierre (2019). "Microeconometric models of consumer demand". In: *Handbook of the Economics of Marketing*. Vol. 1. Elsevier, pp. 1–68.

Dudík, Miroslav et al. (2014). "Doubly robust policy evaluation and optimization". In.

Dunn, Halbert L (1946). "Record linkage". In: *American Journal of Public Health and the Nations Health* 36.12, pp. 1412–1416.

Dwork, Cynthia, Krishnaram Kenthapadi, et al. (2006). "Our data, ourselves: Privacy via distributed noise generation". In: *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, pp. 486–503.

Dwork, Cynthia, Frank McSherry, et al. (2006). "Calibrating noise to sensitivity in private data analysis". In: *Theory of cryptography conference*. Springer, pp. 265–284.

Evans, Georgina et al. (2019). "Statistically valid inferences from privacy protected data". In: *American Political Science Review*.

Fader, Peter S, Bruce GS Hardie, and Ka Lok Lee (2005). "RFM and CLV: Using iso-value curves for customer base analysis". In: *Journal of Marketing Research* 42.4, pp. 415–430.

Feit, Elea M and Eric T Bradlow (2021). "Handbook of Market Research". In: Springer, Cham. Chap. Fusion Modeling.

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

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

Fung, Benjamin CM et al. (2010). "Privacy-preserving data publishing: A survey of recent developments". In: *ACM Computing Surveys (Csur)* 42.4, pp. 1–53.

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

Gómez-Bombarelli, Rafael et al. (2018). “Automatic chemical design using a data-driven continuous representation of molecules”. In: *ACS central science* 4.2, pp. 268–276.

Graham, John W (2009). “Missing data analysis: Making it work in the real world”. In: *Annual review of psychology* 60, pp. 549–576.

Griffiths, Ryan-Rhys and José Miguel Hernández-Lobato (2020). “Constrained Bayesian optimization for automatic chemical design using variational autoencoders”. In: *Chemical science* 11.2, pp. 577–586.

Gronau, Reuben (1977). “Leisure, home production, and work–the theory of the allocation of time revisited”. In: *Journal of political economy* 85.6, pp. 1099–1123.

Groves, Robert M et al. (2011). *Survey methodology*. John Wiley & Sons.

Gu, Mengyang, Debarun Bhattacharjya, and Dharmashankar Subramanian (2018). “Nonparametric Estimation of Utility Functions”. In: *ArXiv Preprint arXiv:1807.10840*.

Gustafsson, Anders, Michael D Johnson, and Inger Roos (2005). “The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention”. In: *Journal of Marketing* 69.4, pp. 210–218.

Hanemann, W Michael (1984). “Discrete/Continuous Models of Consumer Demand”. In: *Econometrica: Journal of the Econometric Society*, pp. 541–561.

Hasegawa, Shohei, Nobuhiko Terui, and Greg M Allenby (2012). “Dynamic brand satiation”. In: *Journal of Marketing Research* 49.6, pp. 842–853.

Hoffman, Matthew D and Andrew Gelman (2014). “The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.” In: *J. Mach. Learn. Res.* 15.1, pp. 1593–1623.

Jackman, Simon (2001). “Multidimensional analysis of roll call data via Bayesian simulation: Identification, estimation, inference, and model checking”. In: *Political Analysis* 9.3, pp. 227–241.

Jälkö, Joonas, Onur Dikmen, and Antti Honkela (2017). “Differentially Private Variational Inference for Non-conjugate Models”. In: *Uncertainty in Artificial Intelligence 2017*.

Kaissis, Georgios et al. (2021). “End-to-end privacy preserving deep learning on multi-institutional medical imaging”. In: *Nature Machine Intelligence* 3.6, pp. 473–484.

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

— (2000). “Factor analysis and missing data”. In: *Journal of Marketing Research* 37.4, pp. 490–498.

Kannan, PK and Gordon P Wright (1991). “Modeling and testing structured markets: A nested logit approach”. In: *Marketing Science* 10.1, pp. 58–82.

Kapteyn, Michael G, Jacob VR Pretorius, and Karen E Willcox (2021). “A probabilistic graphical model foundation for enabling predictive digital twins at scale”. In: *Nature Computational Science* 1.5, pp. 337–347.

Keiningham, Timothy L et al. (2007). "A longitudinal examination of net promoter and firm revenue growth". In: *Journal of Marketing* 71.3, pp. 39–51.

Kekre, Sunder, Mayuram S Krishnan, and Kannan Srinivasan (1995). "Drivers of customer satisfaction for software products: implications for design and service support". In: *Management Science* 41.9, pp. 1456–1470.

Kim, Chul et al. (2023). "Outside good utility and substitution patterns in direct utility models". In: *Journal of choice modelling* 49, p. 100447.

Kim, Jaehwan, Greg M Allenby, and Peter E Rossi (2002). "Modeling Consumer Demand for Variety". In: *Marketing Science* 21.3, pp. 229–250.

Kim, Kwangjo and Harry Chandra Tanuwidjaja (2021). *Privacy-preserving Deep Learning: A Comprehensive Survey*. Springer.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kingma, Diederik P and Max Welling (2013). "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114*.

— (2019). "An introduction to variational autoencoders". In: *arXiv preprint arXiv:1906.02691*.

Klami, Arto, Seppo Virtanen, and Samuel Kaski (2013). "Bayesian Canonical Correlation Analysis." In: *Journal of Machine Learning Research* 14.4.

Kobyzev, Ivan, Simon JD Prince, and Marcus A Brubaker (2020). "Normalizing flows: An introduction and review of current methods". In: *IEEE transactions on pattern analysis and machine intelligence* 43.11, pp. 3964–3979.

Lee, Thomas Y and Eric T Bradlow (2011). "Automated marketing research using online customer reviews". In: *Journal of Marketing Research* 48.5, pp. 881–894.

Lemmens, Aurélie and Christophe Croux (2006). "Bagging and boosting classification trees to predict churn". In: *Journal of Marketing Research* 43.2, pp. 276–286.

Lemmens, Aurélie and Sunil Gupta (2020). "Managing churn to maximize profits". In: *Marketing Science* 39.5, pp. 956–973.

Letham, Benjamin and Eytan Bakshy (2019). "Bayesian optimization for Policy Search via Online-Offline Experimentation." In: *Journal of Machine Learning Research* 20, pp. 145–1.

Li, Lucy et al. (2019). "Protocols for checking compromised credentials". In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1387–1403.

Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian (2007). "t-closeness: Privacy beyond k-anonymity and l-diversity". In: *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, pp. 106–115.

Li, Shaobo et al. (2022). "Reidentification Risk in Panel Data: Protecting for k-Anonymity". In: *Information Systems Research*.

Lin, Tesary (2022). "Valuing intrinsic and instrumental preferences for privacy". In: *Marketing Science* 41.4, pp. 663–681.

Lin, Tesary and Sanjog Misra (2022). "Frontiers: the identity fragmentation bias". In: *Marketing Science* 41.3, pp. 433–440.

Little, Roderick JA and Donald B Rubin (1989). "The analysis of social science data with missing values". In: *Sociological Methods & Research* 18.2-3, pp. 292–326.

— (2019). *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons.

Liu, Jia et al. (2020). "Urban big data fusion based on deep learning: An overview". In: *Information Fusion* 53, pp. 123–133.

Machanavajjhala, Ashwin et al. (2007). "l-diversity: Privacy beyond k-anonymity". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1, 3–es.

Maler, Karl-Goran (1974). *Environmental economics: a Theoretical Inquiry*. Routledge.

Malshe, Ashwin, Anatoli Colicev, and Vikas Mittal (2020). "How main street drives wall street: Customer (dis) satisfaction, short sellers, and abnormal returns". In: *Journal of Marketing Research* 57.6, pp. 1055–1075.

Mason, Karen Oppenheim et al. (1973). "Some methodological issues in cohort analysis of archival data". In: *American sociological review*, pp. 242–258.

McCarthy, Daniel Minh and Elliot Shin Oblander (2021). "Scalable Data Fusion with Selection Correction: An Application to Customer Base Analysis". In: *Marketing Science*.

Narayanan, Arvind and Vitaly Shmatikov (2008). "Robust de-anonymization of large sparse datasets". In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, pp. 111–125.

Narra, Krishna Giri and Chiyuan Zhang (2022). *Private ads prediction with dp-sgd*. en. URL: https://ai.googleblog.com/2022/12/private-ads-prediction-with-dp-sgd.html?m=1.

Neslin, Scott A et al. (2006). "Defection detection: Measuring and understanding the predictive accuracy of customer churn models". In: *Journal of marketing research* 43.2, pp. 204–211.

Netzer, Oded et al. (2012). "Mine your own business: Market-structure surveillance through text mining". In: *Marketing Science* 31.3, pp. 521–543.

Papamakarios, George et al. (2021). "Normalizing Flows for Probabilistic Modeling and Inference." In: *J. Mach. Learn. Res.* 22.57, pp. 1–64.

Papernot, Nicolas (2019). "Machine Learning at Scale with Differential Privacy in TensorFlow". In: *2019 {USENIX} Conference on Privacy Engineering Practice and Respect ({PEPR} 19)*, pp. 1–1.

Phan, Du, Neeraj Pradhan, and Martin Jankowiak (2019). "Composable effects for flexible and accelerated probabilistic programming in NumPyro". In: *arXiv preprint arXiv:1912.11554*.

Ping, Haoyue, Julia Stoyanovich, and Bill Howe (2017). "Datasynthesizer: Privacy-preserving synthetic datasets". In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, pp. 1–5.

Pinjari, Abdul Rawoof and Chandra Bhat (2021). "Computationally efficient forecasting procedures for Kuhn-Tucker consumer demand model systems: Application to residential energy consumption analysis". In: *Journal of choice modelling* 39, p. 100283.

Pollak, Robert A and Michael L Wachter (1975). "The relevance of the household production function and its implications for the allocation of time". In: *Journal of Political Economy* 83.2, pp. 255–278.

Prediger, Lukas et al. (2022). "d3p-A Python Package for Differentially-Private Probabilistic Programming". In: *Proceedings on Privacy Enhancing Technologies* 2, pp. 407–425.

Qian, Yi and Hui Xie (2014). "Which brand purchasers are lost to counterfeiters? An application of new data fusion approaches". In: *Marketing Science* 33.3, pp. 437–448.

— (2022). "Simplifying Bias Correction for Selective Sampling: A Unified Distribution-Free Approach to Handling Endogenously Selected Samples". In: *Marketing Science* 41.2, pp. 211–432.

Rasmussen, Carl Edward, Christopher KI Williams, et al. (2006). *Gaussian processes for machine learning*. Vol. 1. Springer.

Reichheld, Fred (2011). *The ultimate question 2.0 (revised and expanded edition): How net promoter companies thrive in a customer-driven world*. Harvard Business Review Press.

Reichheld, Frederick F (2003). "The one number you need to grow". In: *Harvard Business Review* 81.12, pp. 46–55.

Rezende, Danilo and Shakir Mohamed (2015). "Variational inference with normalizing flows". In: *International Conference on Machine Learning*. PMLR, pp. 1530–1538.

Rezende, Danilo Jimenez and Fabio Viola (2018). "Taming vaes". In: *arXiv preprint arXiv:1810.00597*.

Riihimäki, Jaakko and Aki Vehtari (2010). "Gaussian processes with monotonicity information". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, pp. 645–652.

Ringel, Daniel M (2023). "Multimarket membership mapping". In: *Journal of Marketing Research* 60.2, pp. 237–262.

Riutort-Mayol, Gabriel et al. (2023). "Practical Hilbert space approximate Bayesian Gaussian processes for probabilistic programming". In: *Statistics and Computing* 33.1, p. 17.

Rossi, Peter E and Greg M Allenby (2003). "Bayesian statistics and marketing". In: *Marketing Science* 22.3, pp. 304–328.

Rubin, Donald B (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of Educational Psychology* 66.5, p. 688.

Rubin, Donald B (1976). “Inference and missing data”. In: *Biometrika* 63.3, pp. 581–592.

Ryffel, Theo et al. (2018). “A generic framework for privacy preserving deep learning”. In: *arXiv preprint arXiv:1811.04017*.

Schoenmueller, Verena, Oded Netzer, and Florian Stahl (2020). “The polarity of online reviews: Prevalence, drivers and implications”. In: *Journal of Marketing Research* 57.5, pp. 853–877.

Shokri, Reza and Vitaly Shmatikov (2015). “Privacy-preserving deep learning”. In: *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321.

Solin, Arno and Simo Särkkä (2020). “Hilbert space methods for reduced-rank Gaussian process regression”. In: *Statistics and Computing* 30.2, pp. 419–446.

Swait, Joffre and Rick L Andrews (2003). “Enriching scanner panel models with choice experiments”. In: *Marketing Science* 22.4, pp. 442–460.

Sweeney, Latanya (1997). “Weaving technology and policy together to maintain confidentiality”. In: *The Journal of Law, Medicine & Ethics* 25.2-3, pp. 98–110.

— (2002). “k-anonymity: A model for protecting privacy”. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05, pp. 557–570.

Takagi, Shun et al. (2020). “P3GM: Private High-Dimensional Data Release via Privacy Preserving Phased Generative Model”. In: *arXiv preprint arXiv:2006.12101*.

Tirunillai, Seshadri and Gerard J Tellis (2014). “Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation”. In: *Journal of Marketing Research* 51.4, pp. 463–479.

Tsialiamanis, George et al. (2021). “On generative models as the basis for digital twins”. In: *Data-Centric Engineering* 2, e11.

U.S. Census Bureau (2021). *Disclosure avoidance for the 2020 census: An introduction*. URL: https://www2.census.gov/library/publications/decennial/2020/2020-census-disclosure-avoidance-handbook.pdf.

Unger, Moshe et al. (2018). “Inferring contextual preferences using deep encoder-decoder learners”. In: *New Review of Hypermedia and Multimedia* 24.3, pp. 262–290.

Van Soest, Arthur, Arie Kapteyn, and Peter Kooreman (1993). “Coherency and regularity of demand systems with equality and inequality constraints”. In: *Journal of Econometrics* 57.1-3, pp. 161–188.

Venkatesan, Rajkumar and Vita Kumar (2004). “A customer lifetime value framework for customer selection and resource allocation strategy”. In: *Journal of Marketing* 68.4, pp. 106–125.

Wales, Terence J and Alan Donald Woodland (1983). “Estimation of Consumer Demand Systems with Binding Non-Negativity Constraints”. In: *Journal of Econometrics* 21.3, pp. 263–285.

Wang, Xiaojing and James O Berger (2016). "Estimating Shape Constrained Functions Using Gaussian Processes". In: *SIAM/ASA Journal on Uncertainty Quantification* 4.1, pp. 1–25.

Yaari, Menahem E (1977). "A note on separability and quasiconcavity". In: *Econometrica: Journal of the Econometric Society*, pp. 1183–1186.

Yang, Yang (2006). "Bayesian inference for hierarchical age-period-cohort models of repeated cross-section survey data". In: *Sociological Methodology* 36.1, pp. 39–74.

Yu, Jinsong et al. (2021). "A Digital Twin approach based on nonparametric Bayesian network for complex system health monitoring". In: *Journal of Manufacturing Systems* 58, pp. 293–304.

Zhang, Cheng et al. (2018). "Advances in variational inference". In: *IEEE transactions on pattern analysis and machine intelligence* 41.8, pp. 2008–2026.

Zhang, Hao (2004). "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics". In: *Journal of the American Statistical Association* 99.465, pp. 250–261.

Zheng, Yu (2015). "Methodologies for cross-domain data fusion: An overview". In: *IEEE transactions on big data* 1.1, pp. 16–34.