Probabilistic Machine Learning: New Frontiers for Modeling Consumers and their Choices*

Ryan Dew, Nicolas Padilla, Lan E. Luo, Shin Oblander,
Asim Ansari, Khaled Boughanmi, Michael Braun, Fred Feinberg,
Jia Liu, Thomas Otter, Longxiu Tian, Yixin Wang, Mingzhang Yin†

April 30, 2024

Abstract

Making sense of massive, individual-level data is challenging: marketing researchers and analysts need flexible models that can accommodate rich patterns of heterogeneity and dynamics, work with and link diverse data types, and scale to modern data sizes. Practitioners also need tools that can quantify uncertainty in models and predictions of consumer behavior to inform optimal decision-making. In this paper, we demonstrate the promise of probabilistic machine learning (PML), which refers to the pairing of probabilistic modeling and machine learning methods, in pushing the frontier of combining flexibility, scalability, interpretability, and uncertainty quantification for building better models of consumers and their choices. Specifically, we overview both PML models and inference methods, and highlight their utility for addressing four common classes of marketing problems: (1) uncovering heterogeneity, (2) flexibly modeling nonlinearities and dynamics, (3) handling high-dimensional and unstructured data, and (4) addressing missingness, often via data fusion. We also discuss promising directions in enriching marketing models, reflecting recent developments in representation learning, causal inference, experimentation and decision-making, and theory-based behavioral modeling.

Keywords: machine learning, Bayesian statistics, Bayesian nonparametrics, generative models, unstructured data, representation learning, causal inference

*This paper is based on a session from the 2023 Triennial Choice Symposium. Ryan Dew and Nicolas Padilla organized the session. Lan Luo and Shin Oblander led the writing and editing process. All other authors contributed equally and are listed alphabetically. Ryan Dew is the corresponding author: ryandew@wharton.upenn.edu.

†Affiliations: Ryan Dew, University of Pennsylvania; Nicolas Padilla, London Business School; Lan Luo, Shin Oblander, and Asim Ansari, Columbia University; Khaled Boughanmi, Cornell University; Michael Braun, Southern Methodist University; Fred Feinberg and Yixin Wang, University of Michigan; Jia Liu, Hong Kong University of Science and Technology; Thomas Otter, Goethe Universität Frankfurt and NOVA SBE, Lisbon; Longxiu Tian, University of North Carolina, Chapel Hill; Mingzhang Yin, University of Florida.

1 Introduction

Modern marketing contexts involve large datasets characterized by rich patterns of individual-level variation and dynamics. Such data is often of different types, such as structured tabular data and unstructured data involving text, images, and network relations. Flexible and expressive models are needed to capture complex variation in data, and scalable computational methods are required to infer their numerous latent variables. Moreover, proper uncertainty quantification is necessary for prediction and for making optimal decisions. In this paper, we describe how probabilistic machine learning (PML), or the pairing of probabilistic modeling and machine learning methods, can deliver this combination of flexibility, scalability, and uncertainty quantification to model marketing phenomena.

To better understand PML, and its promise for marketing research, we first define its two ingredients: probability modeling and machine learning. Probability models specify the statistical relationships among variables, observed or latent, using probability distributions. While this definition can include almost any statistical or econometric model, probability models are differentiated by jointly modeling the data generating process (DGP) of both observed and latent variables. These latent variables (or parameters) capture the structure behind how the data was generated. Their values are commonly estimated using Bayesian or quasi-Bayesian methods, which yield properly calibrated estimates of uncertainty of the parameters, even in small samples. While probability models are popular in marketing, they often: (1) make strong parametric assumptions about data generating processes, which may not reflect reality; and (2) rely on slow and computationally costly methods like Markov chain Monte Carlo (MCMC) for inference, which cannot scale to modern data sizes.

Machine learning (ML) methods, in contrast, deliver flexibility and scalability, often at the expense of interpretability and uncertainty quantification. ML most commonly refers to flexible models (e.g., neural networks), regularization methods (e.g., LASSO), and scalable computational techniques (e.g., stochastic gradient descent and back-propagation) for dealing with large and complex data (Goodfellow et al. 2016). However, traditional ML methods often do not have clearly interpretable outputs (Rudin 2019) or adequately quantify uncertainty, and are therefore not ideal for decision-making, especially in high risk situations.

PML integrates these two approaches by using ML methods in tandem with probability models (Murphy 2022). This unification uses probability models to represent latent structures, but unlike traditional approaches, draws on advances in ML to flexibly specify these probability distributions. This enables the modeling of complex patterns in large datasets with minimal assumptions and facilitates scalable inference. The combination of a principled approach to statistical reasoning afforded by probabilistic models and the flexibility and scalability afforded by ML makes PML promising for developing marketing models. Indeed, recent research in marketing is realizing this potential: throughout the rest of this paper, we highlight how PML models and ideas have already influenced, and will continue to influence, how researchers represent marketing phenomena, discover patterns, generate and synthesize data (e.g., *unstructured* data like text and images), make predictions, and decide optimal actions in a data-driven manner.

Motivating Example: Choice Models To make these ideas more concrete, we now discuss them within the classic marketing context of discrete choice models. Consider data on choices y_{it} made by a large number of consumers $i = 1, \ldots, N$ over repeated purchase occasions $t = 1, \ldots, T_i$. The consumers are assumed to choose from a choice set of alternatives $j = 1, \ldots, J$. The choice alternatives are described by time varying attributes and the consumers are represented by their characteristics. These two sets of variables can be collected in a vector $\mathbf{x}_{ijt}$. The goals of choice modeling are to understand consumer preferences, characterize the heterogeneity in preferences across consumers, and predict future choices.

A simple model such as the multinomial logit can be used to capture the probabilities of choosing the different alternatives. In a typical logit model, a parametric, linear utility function $u_{ijt} = \boldsymbol{\theta}'\mathbf{x}_{ijt} + \epsilon_{ijt}$, is used to specify consumer i's utility for alternative j. Assuming ϵ_{ijt} is IID Gumbel, this yields standard logit choice probabilities:

$$p\left(Y_{it} = j \mid \{\mathbf{x}_{ikt}\}_{k=1}^{J}, \boldsymbol{\theta}\right) = \frac{e^{\boldsymbol{\theta}'\mathbf{x}_{ijt}}}{\sum_k e^{\boldsymbol{\theta}'\mathbf{x}_{ikt}}}. \tag{1}$$

The researcher must then find values of the parameters $\boldsymbol{\theta}$ based on the data $\mathcal{D} = \left\{\mathbf{x}_{ijt}, y_{it}\right\}$ so as to accurately characterize the statistical relationship between $\mathbf{x}$ and y. Both probabilistic modeling and standard ML provide approaches for doing so.

The probabilistic Bayesian approach combines this model with a prior distribution on $\boldsymbol{\theta}$ to obtain a full DGP, from which we can compute the posterior distribution of $\boldsymbol{\theta}$, $p(\boldsymbol{\theta} \mid \mathcal{D})$. This model specification can be extended to allow, for instance, unobserved heterogeneity reflected by individual-level coefficients $\boldsymbol{\theta}_i$, which may share a common prior. This results in a mixed logit specification, in which consumer preferences are modeled as latent variables drawn from a common mixing distribution. Assuming a specific distribution for $\boldsymbol{\theta}_i$ (e.g., $\boldsymbol{\theta}_i \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$) adds structure to the problem, and estimating $\boldsymbol{\mu}$ and Σ jointly with $\boldsymbol{\theta}_i$ allows for statistical information to be shared across individuals, resulting in more efficient inference. The challenge of this approach is inference: to compute estimates of individual-level preferences and the uncertainty around them, we need to evaluate a potentially intractable posterior distribution for every individual. Often, this is done by slow and computationally burdensome sampling techniques like MCMC, which scale poorly to large datasets with many individuals.

In an ML approach, the model in Equation 1 is an example of supervised classification. Its probability expression is referred to as the softmax function, and the optimal $\boldsymbol{\theta}$ is learned by minimizing a loss function based on that probability. This formulation offers more flexibility: for example, one can replace the linear utilities with a generic ("black box") function, such as a deep neural network. Such an approach allows allow for nonlinearities and interaction effects in how the observed features relate to the predicted choice probabilities, and can accommodate complex, high-dimensional x's like product images, text, and audio. Modern computational techniques like stochastic optimization and automatic differentiation allow such a model to be estimated at scale on massive datasets. At the same time, its black box nature means that the statistical relationships learned between variables are difficult to interpret. The model may not capture important structures in the data such as hierarchies, dynamics, and patterns implied by microeconomic axioms (e.g., downward-sloping demand), and it can become difficult to incorporate unobserved heterogeneity in choice behavior across consumers and quantify statistical uncertainty.

A PML approach can unite the strengths of both perspectives, enabling scalable yet interpretable inferences about consumer preferences, while maintaining uncertainty quantification. For instance, the mixed logit model can be estimated using variational inference (VI), which approximates the posterior of individual-level parameters with a simpler, more tractable distribution, replacing costly MCMC iterations with an optimization routine (Braun and McAuliffe 2010).

This optimization-based formulation facilitates ML computational techniques like automatic differentiation and stochastic gradients, enabling scalability. PML methods can also extend choice models to incorporate data that is large across other dimensions, such as large choice sets across many product categories (Ruiz et al. 2020) or large attribute spaces, where products are described in terms of unstructured data like images (Dew 2023). Importantly, unlike in typical ML, these models are built upon Bayesian statistics, providing a natural way to reason about model uncertainty. That being said, there are tradeoffs: for instance, accurately approximating the posterior often comes at the cost of scalability, especially when models are highly complex. Crucially, PML enables researchers to choose exactly which of these tradeoffs they are willing to make.

The benefits of the PML approach have direct implications for not only understanding choice but also for making decisions. Using models like Equation 1 for decisions like who to target with an ad or coupon has a long history in marketing (e.g., Rossi et al. 1996). Decisions like this remain relevant for modern online platforms, but at much larger scales and lower latencies: given a user's demographics and history, platforms must determine the optimal product to recommend or advertisement to deliver, often within milliseconds. PML provides a path forward for optimal decision-making in such contexts. By combining PML architectures with Bayesian decision theory, we can derive optimal policies that properly integrate uncertainty *and* feasibly scale up to modern settings involving large, complex data. In situations where the posterior is well-approximated, such a decision-theoretic approach is superior to ML approaches that rely on plug-in estimates of model parameters, which can lead to overly risky decisions and misestimated profits, especially for decisions based on few observations.

Roadmap As outlined above, the probabilistic approach to machine learning offers many potential benefits. Hereafter, we describe this approach in more detail, highlighting both how it has already been used in marketing and choice contexts (whether explicitly labeled as PML or not) and where we see the opportunities for how PML can help push the field forward. Specifically, in Section 2, we briefly review the foundations of PML and discuss directly how PML can inform managerial decision-making. In Section 3, we discuss key modeling ideas, organized around common problems addressed by marketers. In Section 4, we delineate inference procedures used in PML, with a focus on scalability and practicality. In Section 5, we highlight interesting directions

for future research in applying PML to marketing problems. We conclude in Section 6.

2 Bayes is Dead, Long Live Bayes: Foundations of PML

2.1 Defining Models with Probabilities

The core of PML involves defining a joint probability distribution over data $\mathcal{D}$, and the latent variables $\boldsymbol{\theta}$ that govern the data generating process. The joint distribution can be written as $p(\mathcal{D}, \boldsymbol{\theta}) = p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$; $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where, $p(\mathcal{D} \mid \boldsymbol{\theta})$ is the data likelihood that describes how the data are generated conditional on the latent variables, and $p(\boldsymbol{\theta})$ is the prior distribution that specifies how $\boldsymbol{\theta}$ is generated. The prior captures our beliefs about the latent variables, absent any data, allowing incorporation of prior substantive knowledge and enabling better extrapolation from limited observations (analogous to inductive biases in ML). Well-chosen priors often yield favorable statistical properties of model estimates, effectively regularizing models in interpretable ways. To visually summarize probability models, researchers often use directed acyclic graphs (DAGs; Pearl 1988; Koller and Friedman 2009). DAGs make explicit the conditional dependence structure between observed and latent variables, clarifying the flow of the data generating process. An example DAG for the mixed logit model from Section 1 is given in Figure 1.

Given this framework, a natural question to ask is, given the data, what are the likely values of the latent variables? To answer this question, we can derive the posterior distribution of the latent variables using Bayes' theorem:

$$p\left(\boldsymbol{\theta} \mid \mathcal{D}\right) = \frac{p\left(\mathcal{D}, \boldsymbol{\theta}\right)}{p\left(\mathcal{D}\right)} = \frac{p\left(\mathcal{D} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)}{\int_{\boldsymbol{\Theta}} p\left(\mathcal{D} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}} \tag{2}$$

We can also reason about likely values of new data, through the posterior predictive distribution,

$$p\left(\mathcal{D}_{\text{new}} \mid \mathcal{D}\right) = \int_{\boldsymbol{\Theta}} p\left(\mathcal{D}_{\text{new}} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \mathcal{D}\right) d\boldsymbol{\theta}, \tag{3}$$

which represents the uncertainty about future values of the data, conditional on the current data and model. Putting these expressions in context, the posterior distribution (Equation 2) tells us about a parameter of interest, like price sensitivity, while the posterior predictive distribution

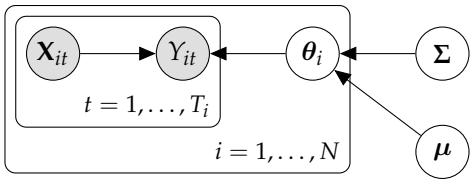

Figure 1: DAG for the mixed logit model.
In the graph, each node is a variable. Shaded nodes are observed, while unshaded nodes are latent. Arrows represent which variables' distributions depend on which other variables. Boxes, typically referred to as "plates," capture indices, representing IID draws. For example, θ_i is drawn IID for $i = 1, \ldots, N$, conditional on μ and Σ, and all of these variables are latent.

(Equation 3) is used to predict future consumer behavior, e.g., whether a customer will make a purchase in the next time period. These probability expressions are valid regardless of sample size: a key benefit of this perspective is that we can in principle perform exact inference, even in finite samples. Accurate finite sample uncertainty quantification is crucial in applications like individual targeting (as in Section 1), where there will often only be a few observations per individual.

2.2 Connections to Machine Learning

In contrast to the Bayesian perspective, ML models typically do not formally distinguish between the likelihood and the prior distribution. Nevertheless, there are deep links between probability models and machine learning models, which help us understand how PML integrates both approaches. To illustrate these connections, consider computing a point estimate of the latent variables under the modeling framework described above. Bypassing, for now, the more difficult task of deriving the full posterior of $\boldsymbol{\theta}$, we can obtain its point estimate by maximizing the log of Equation 2. This yields what is called the maximum a posteriori or "MAP" estimate,

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \log p\left(\boldsymbol{\theta} \mid \mathcal{D}\right) = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left[\log p\left(\mathcal{D} \mid \boldsymbol{\theta}\right) + \log p\left(\boldsymbol{\theta}\right)\right], \tag{4}$$

which, intuitively, gives the mode of the posterior distribution. This objective can be interpreted as a penalized likelihood, where the penalty term is equivalent to the log-prior.

In ML, such objective functions are common: the term that depends on the data is called the loss, and the term that depends just on the value of the parameters is called the regularizer. Typically, ML models pose the problem as a minimization, rather than maximization. Consequently,

a probability model implies a loss function equal to its negative log-likelihood (NLL). Interestingly, many commonly used loss functions in ML correspond to NLLs of probability models. For example, the cross-entropy loss used in supervised classification is equivalent to the NLL of a categorical distribution. The mean squared error loss is equivalent to the NLL of a Gaussian distribution. Thus, choosing a specific loss function can be viewed as taking a stance on the (conditional) distribution of the observables.

Additionally, this analogy allows us to see the role of the prior as a regularizer. In particular, commonly used regularizers in ML can be interpreted as imposing specific prior distributions on model parameters. For instance, for a uniform prior, the $\log p(\boldsymbol{\theta})$ term becomes a constant, and MAP estimation reduces simply to maximum likelihood without any regularization. Likewise, specific models such as ridge (LASSO) regression can be thought of as a Bayesian linear regression model with a Gaussian (Laplace) prior (Murphy 2022), while a neural network with weight decay (i.e., an L2 penalty on the weights) can be thought of as a Bayesian neural network with Gaussian priors (Neal 2012). These equivalences imply that many commonly used ML models can be interpreted as implicitly estimating probabilistic models under specific assumptions about the distribution of the observables and the priors. Explicitly articulating these assumptions and applying probabilistic reasoning can help make such models more interpretable and elucidate opportunities to relax distributional assumptions. Furthermore, expressing models probabilistically allows us to use Bayesian inference to reason about uncertainty in latent variables.

2.3 Computation

Often, computation is a barrier to implementing probability models in practice. Except in special cases, the denominator of the posterior Equation 2, also called the marginal likelihood or normalizing constant, cannot be solved in closed-form. As a result, the posterior density is only known up to proportionality: $p(\boldsymbol{\theta} \mid \mathcal{D}) \propto p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$. Foundational research in Bayesian statistics and econometrics has derived ways to bypass this barrier or approximate the posterior. Perhaps the most widely adopted solution is Markov chain Monte Carlo (MCMC), which involves designing Markov chains that can be used to draw from the posterior to then infer quantities of interest. However, most early MCMC methods either involve challenging model-specific derivations or

scale poorly, which has inhibited the adoption of Bayesian methods.

To adapt Bayesian machinery to real-world problems, the PML field has developed both improvements to MCMC methods, as well as alternative methods for approximating Equation 2, enabling Bayesian computation to be accessible even at scale. For example, Hamiltonian Monte Carlo (Duane et al. 1987; Neal 2011) uses gradient information to improve the efficiency of classic MCMC methods, and can be implemented without any model-specific derivations via automatic differentiation methods. As an alternative to sampling methods, variational inference (Jordan et al. 1999; Blei et al. 2017) instead approximates Equation 2 using a simpler family of distributions, posing the inference problem as an optimization problem, in which the objective is to make the approximation as close as possible to the posterior.[1] This optimization can be augmented with techniques from ML, including automatic differentiation, stochastic optimization, and mini-batching to further improve scalability. We discuss these methods in greater detail in Section 4. These new methods help not just with implementation, but also with model development: since they are both generic—meaning they can be used for any model, without extensive model-specific coding or derivations—and fast, they enable researchers to not only build models, but also assess their performance and easily *modify* the models based on that assessment.[2]

2.4 Model-based Decision-making

Thus far, we have overviewed the ingredients for building probability models and inferring their latent variables in scalable and practical ways. Still, why should marketers care? In particular, how does PML connect to making good marketing decisions? To understand the benefits of PML for decision-making, we first need to formalize what it means to make a data-driven decision. For that, we turn to decision theory: given a model with latent variables $\boldsymbol{\theta}$, we define the decision loss from an action $a \in \mathcal{A}$ as $\mathcal{L}(a, \boldsymbol{\theta})$.[3] This decision loss reflects the objective of the decision-maker: in the choice modeling example from Section 1, the profit maximization objective of the firm translates into a loss that equals (negative) expected profit. The profit depends on consumer

[1] Note a difference in parlance across fields: in PML, the process of computing the posterior is often called "inference," hence the name "variational inference." Such a procedure can then be used to compute quantities like point estimates and credible intervals, which is more aligned with the statistical and econometric usage of the term inference.

[2] The cycle of building models, inferring their latent variables, critiquing their assumptions, and then improving them is referred to as "Box's Loop" by Blei (2014), adapting classic ideas from George Box and collaborators (e.g., Box and Hunter 1962). Efficient, model-agnostic inference methods enables faster iteration through this loop.

[3] We use the term "decision loss" to differentiate from the loss function terminology used in ML.

choices that are a function of preferences $\boldsymbol{\theta}$ and the action taken by the firm, a.[4] In this setup, the Bayes optimal action minimizes the posterior expected loss:

$$\arg\min_{a \in \mathcal{A}} \mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{L}\left(a, \boldsymbol{\theta}\right) \mid \mathcal{D}\right] = \arg\min_{a \in \mathcal{A}} \int_{\boldsymbol{\Theta}} \mathcal{L}\left(a, \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \mathcal{D}\right) d\boldsymbol{\theta} \tag{5}$$

This is the best action *after taking into account model uncertainty*, since Equation 5 depends on the posterior of $\boldsymbol{\theta}$, and thus requires the use of Bayesian inference, as in PML.

To see why the Bayes optimal action is a compelling solution to the decision problem, it is helpful to contrast it to alternatives. For example, rather than integrating over uncertainty in the latent variables, one could simply plug a point estimate $\hat{\boldsymbol{\theta}}$ into the decision loss function, evaluating and optimizing the decision loss $\mathcal{L}(a, \hat{\boldsymbol{\theta}})$. Because this objective function does not reflect posterior uncertainty in the model parameters, it may lead to selecting risky actions that have low expected loss under the point estimate but high expected loss under other plausible values of $\boldsymbol{\theta}$. This concern applies even to recent work in econometrics, which use flexible ML models to uncover heterogeneous treatment effects (Athey and Imbens 2019; Farrell et al. 2020), which are then used to estimate optimal policy assignments (Athey and Wager 2021). These methods typically make use of asymptotic approximations to quantify uncertainty, which may be inaccurate in cases involving few observations, as in many targeting scenarios. The risks of ignoring uncertainty are further compounded when the plug-in estimates themselves yield an overconfident predictive distribution, as is common in modern neural networks (Guo et al. 2017). PML, combined with the Bayesian decision theoretic approach described above, has the potential to resolve these issues by accounting for uncertainty in both the inherent noise in the data (i.e., the error term in the predictive distribution) *and* the model parameters (i.e., posterior uncertainty through $p\left(\boldsymbol{\theta} \mid \mathcal{D}\right)$). Moreover, by integrating elements of machine learning models *within* the Bayesian framework, PML models are often more flexible and accurate than classic models, thus forming an even stronger foundation for decisions.

[4] The profit is *expected* since realized profit is not directly a function of $\boldsymbol{\theta}$ and a, but rather a function of future consumer choices. The decision loss as defined here integrates over the predictive distribution of consumer choices conditional on preference parameters and firm action.

2.5 No Free Lunch: The Iron Simplex

PML and Bayesian decision theory provide a principled framework for making decisions under model uncertainty with limited observations. However, they require well-specified models and *accurate approximations* to the intractable posterior distribution to work well. In specifying PML models, there is often a trade-off: accurately capturing uncertainty often comes at the expense of scalability (e.g., by using methods like HMC instead of VI for inference). Moreover, the more flexible a model is, the more challenging it can be to perform fast, accurate inference and interpret the model. This trade-off between scalability, accuracy, flexibility, and interpretation is what we refer to as the *Iron Simplex*,[5] reflecting the idea that, in building models and using them for decision-making, researchers often must choose how to weigh each of these factors, and make trade-offs among them. Luckily, PML provides tools for explicitly navigating the iron simplex. A host of inference methods are compatible with any one model, and the inherent modularity of probability models coupled with generic inference methods enables researchers to simplify or expand models as needed, without needing to rederive estimation procedures or asymptotic results. Furthermore, as PML methods and computing technologies develop, the iron simplex will continue to expand, potentially enabling faster inference without sacrificing on any other dimensions.

3 Key Modeling Ideas for Marketing Problems

Having described what PML is, we now turn to discussing how PML has been used for modeling consumers and their choices. We organize this section around four common classes of marketing problems where PML has had significant impact: (1) uncovering heterogeneity, (2) flexibly modeling nonlinearities and dynamics, (3) handling high-dimensional data, and (4) addressing missingness, often via data fusion.

[5] The name "Iron Simplex" derives from the concept of the "iron triangle" in product management, which describes the trade-offs between budget, schedule, and scope in determining project quality. In geometry, a simplex is a generalization of a triangle, and the term is often used in probabilistic modeling to describe non-negative vectors that sum to 1.

3.1 Heterogeneity

As we described in Section 1, understanding consumer heterogeneity is crucial for segmentation and targeting. Here, we begin by describing classic hierarchical models for capturing heterogeneity, then show how advancements in PML—specifically, Bayesian nonparametric methods and mixed membership models—have expanded these models to better characterize heterogeneity, especially *unobserved* heterogeneity.

Hierarchical Models The classic approach for capturing unobserved heterogeneity is through hierarchical models. In hierarchical models, latent variables are organized into groups (e.g., of stores, households, consumers, or physicians), where the variables within a group are exchangeable (i.e., conditionally IID) and governed by a set of higher-order latent variables, which pool statistical information across groups (Betancourt and Girolami 2015). Such models have an extensive history in marketing and choice modeling (Rossi and Allenby 2003), as they provide a natural modeling structure for capturing unobserved heterogeneity between consumers, and for obtaining individual-level estimates of consumer preferences (and uncertainty in those estimates). As an example, consider the heterogeneous extension of the choice model described in Section 1: there, choices are assumed to be independent, conditional on individual-level preference parameters $\boldsymbol{\theta}_i$, which themselves are drawn from a common distribution, like a Gaussian distribution with mean vector $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$. The choices are exchangeable, forming groups at the individual-level, and the $\boldsymbol{\theta}_i$ are exchangeable, forming a group defined by their common prior. In this setup, seeing one customer's purchasing history gives information not only about $\boldsymbol{\theta}_i$, but also about $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, which in turn affects the estimates of $\boldsymbol{\theta}_i$ for other customers. In the PML literature, the parameters that are linked to specific units of observation (e.g., $\boldsymbol{\theta}_i$) are often referred to as local variables, and the common parameters (e.g., $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$), as global variables. As we describe further, many of the latent structures used in PML, especially in the context of choice and marketing, are variants of hierarchical models.

Flexible Models of Distributions Parametric heterogeneity distributions, like $\boldsymbol{\theta}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, are often used in marketing for their simplicity. However, such parametric distributions are restrictive and can potentially mask rich patterns of heterogeneity involving multiple preference

segments (i.e., multimodality), skewness, or heavy tails. They can also result in misleading inferences, such as overestimates of state dependence in choice models (Dubé et al. 2010) or biased estimates of structural parameters in search models (Onzo and Ansari 2024). Alternatively, finite mixture or latent class approaches have been used to specify heterogeneity using discrete mixtures involving a few support points that can be interpreted as consumer segments (Kamakura and Russell 1989). These methods are simple, but restrictive: they assume homogeneity within segments, and that the number of segments is known in advance (Allenby et al. 1998). PML-based innovations in *Bayesian nonparametrics* (BNP) can overcome these limitations: by modeling the mixing distribution as unknown, and giving it a nonparametric prior, these methods allow the data to inform the distribution of heterogeneity directly. Modeling unknown distributions nonparametrically allows for considerable flexibility in capturing many different patterns of heterogeneity, while still retaining the hierarchical model structure that facilitates partial pooling of information across data units.

The Dirichlet process (DP) is the most common mechanism for placing priors on nonparametric distributions. A DP prior has two parameters: a *base distribution* G_0 that is a point of central tendency (intuitively, the "average" distribution; typically, a parametric distribution like a multivariate normal) and a *concentration parameter* $\alpha > 0$ that determines how closely the realizations from the DP resemble the base distribution. Samples from a DP are discrete probability distributions whose support lies within the support of the base distribution G_0. While the DP could be used to model the distribution of observed data, it is more commonly used as a mixing distribution, where the unknown distribution G describes how latent variables are distributed. For example, in a hierarchical model, suppose the conditional distribution of the data Y_{it} is governed by individual-level parameter $\boldsymbol{\theta}_i$; we can write the distribution of $\boldsymbol{\theta}_i$ using a DP:

$$Y_{it} \sim f(\boldsymbol{\theta}_i), \qquad \boldsymbol{\theta}_i \sim G, \qquad G \sim DP(G_0, \alpha). \tag{6}$$

Such a specification is called a Dirichlet Process Mixture (DPM). The DP generates *discrete* distri-

butions (Ferguson 1973), even when G_0 is continuous.[6] When α is small, most probability mass tends to concentrate on a few support points, as in a finite mixture distribution; when α is large, probability mass tends to spread across many support points, resulting in a distribution closer to G_0. This property is useful for tasks like clustering or segmentation, since observations separate into discrete support points without needing to prespecify the number of segments. Often, both α and the parameters of G_0 are inferred from data.

In marketing, DPMs have been widely applied to model heterogeneity. Ansari and Mela (2003) use a DPM to model heterogeneity in user- and email-specific propensities of responding to an email communication. Kim et al. (2004) model heterogeneity in the coefficients of a discrete choice model. Wedel and Zhang (2004) use DPMs to capture store-level heterogeneity in cross-category price effects on aggregated sales. Braun et al. (2006) use DPMs to flexibly model the joint distribution of household-level insurance claim rates and latent deductibles. Braun and Bonfrer (2011) exploit the discreteness of the DP to simplify computation of a social network model in a latent space (discussed further in Section 3.3). Apart from modeling heterogeneity, DPs have also been used in other settings. Ansari and Iyengar (2006) use DPMs to model error distributions in choice models, and Li and Ansari (2014) use centered DPMs that allow for mean and variance restrictions on the unknown distributions to specify the uncertainty on the distribution of structural errors in choice models.

While flexible, the DP still makes several assumptions. For instance, it assumes most probability mass concentrates on a handful of large segments. For applications with a large number of small clusters, the Pitman-Yor process (Pitman and Yor 1997) incorporates an additional parameter to allow for heavy tails, which Padilla et al. (2023) use to account for heterogeneous context-based preferences for online flight searches. Another extension of the DP is the Hierarchical Dirichlet Process (HDP; Teh et al. 2006), which allows grouped data to have separate DP priors per group, linking those priors through a common base distribution, which itself comes from a DP. This structure naturally captures heterogeneity in nested data. For example, Voleti and Ghosh

[6] The "stick-breaking" representation is useful to illustrate this point. A draw $G \sim DP(G_0, \alpha)$ can be represented as an infinite mixture $G(\cdot) = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}(\cdot)$ of discrete points ϕ_k drawn from the base distribution $\phi_k \sim G_0$. The probability mass π_k associated with support point ϕ_k is generated by breaking a stick of initial unit length, recursively, by proportions V_k that are drawn from a Beta distribution $V_k \sim \text{Beta}(1, \alpha)$, yielding probability mass $\pi_k = V_k \cdot \prod_{h<k}(1 - V_h)$. Thus, G will be a mixture of discrete (but infinite) support points. In practice, inference routines either truncate to finitely many support points or perform sampling on the individual-level parameters directly to avoid sampling an infinite-dimensional object.

(2014) model the distribution of SKU-level elasticities in an aggregated demand model using an HDP that leverages brand-SKU hierarchies, and Boughanmi and Ansari (2021) use an HDP to capture the dependency of textual tags of songs that belong to the same album.

Mixed Membership Models The DP is flexible, but fundamentally assumes that consumers can be described by discrete segments. While assigning consumers to discrete segments is intuitively appealing, in many instances, a single segment may not adequately characterize a consumer. For example, one could simultaneously be a fitness enthusiast and gourmet food lover, exhibiting different preferences and behavioral patterns depending on the shopping trip. To capture such phenomena, we can adapt a different tool from PML: mixed membership models. These models add another layer to the modeling hierarchy—instead of assigning a consumer to a single segment, mixed membership models allow consumers to move between segments at different points in time (e.g., on different purchase occasions). Consumers are thus represented not by a single segment assignment, but by discrete distributions that specify the probability of the consumer belonging to each segment for any given observation (Blei 2014). Mixed membership models allow us to retain the interpretability and simplicity of discrete segments while also more flexibly characterizing individuals.[7] Mixed membership specifications have only recently been used to model consumers. Applications include modeling browsing histories in terms of consumer roles (Trusov et al. 2016), understanding purchase histories in terms of projects and motivations (Jacobs et al. 2021; Kim and Zhang 2023), and learning consumer archetypes in marketing research (Kim and Allenby 2022).

Mixed membership models are not limited to modeling heterogeneity. Perhaps the most influential use of MMMs, both in marketing and beyond, has been for topic modeling in text analysis. Topic models characterize textual documents as a mixture of discrete topics based on word usage. The most widely used topic model is latent Dirichlet allocation (LDA, so named for its extensive use of Dirichlet priors; Blei et al. 2003). In LDA, there are K "topics," where a topic is a categorical distribution over a vocabulary of V words, with weights ϕ_k drawn from a V-dimensional, symmetric Dirichlet prior: $\phi_k \sim \text{Dirichlet}_V(\eta)$. Each document is assumed to have a mixed membership over the K topics, with weights $\boldsymbol{\theta}_i$ drawn from a K-dimensional, symmetric Dirichlet prior:

[7] They can also be combined with BNP methods to estimate the appropriate number of segments from the data (e.g., Shi et al. 2023).

$\boldsymbol{\theta}_i \sim \text{Dirichlet}_K(\alpha)$. Each word within a document X_{it} is then modeled as being independently drawn using a two-step procedure, where first a topic Z_{it} is drawn from the document's topic distribution, and then a word is drawn from the topic's word distribution:

$$Z_{it} \mid \boldsymbol{\theta}_i \sim \text{Categorical}_K(\boldsymbol{\theta}_i) \qquad X_{it} \mid Z_{it}, \boldsymbol{\phi}_{Z_{it}} \sim \text{Categorical}_V(\boldsymbol{\phi}_{Z_{it}})$$

Mapping this onto the consumer example above, the topics are the preference segments; documents, the consumers; and words, the individual choices. In marketing, LDA and its extensions have been used to discover semantic themes in textual data involving online reviews (Tirunillai and Tellis 2014; Büschken and Allenby 2016; Puranam et al. 2017), search queries and results (Liu and Toubia 2018), descriptions of entertainment products (Toubia et al. 2019), and social media content (Zhong and Schweidel 2020).

3.2 Nonlinearities and Dynamics

Another way by which PML has enabled better models of consumers is through flexible models for unknown functions. Many analyses can be framed as problems of estimating unknown functions. For instance, a simple regression model linking some outcome Y_i to a set of independent variables $\boldsymbol{X}_i$ is fundamentally a problem of estimating an unknown function, f: $Y_i = f(\boldsymbol{X}_i) + \varepsilon_i$. Similarly, in choice modeling, the utility function links variables $\boldsymbol{X}_i$ to the probability of a consumer choosing a given alternative. While researchers typically assume these functions are linear (possibly with a nonlinear "link function") and static, such assumptions may result in wrong substantive insights and misguided decisions. Consider, for example, marketing mix models: the relationship between marketing spend and revenue may exhibit decreasing marginal effects, non-monotonic patterns, or more complex functional forms that would be masked by assuming linearity. Moreover, the nature of these relationships may change over time as markets evolve. PML methods allow us to avoid making restrictive assumptions through flexible models of unknown functions. These specifications can also be useful in representing dynamics, by considering parameters as functions of time. In this section, we present three PML approaches for estimating unknown functions—Gaussian processes, Bayesian splines, and Bayesian neural networks—and discuss how they have been applied for learning nonlinearities and dynamics in marketing and

choice contexts.

Gaussian processes Gaussian processes (GPs) are used as Bayesian nonparametric priors over unknown functions.[8] Because GPs define a probability distribution on f for *any* set of inputs $\mathbf{x}$, they provide a solution for specifying a prior for an unknown function. The GP specification consists of a mean function, $\mu(\mathbf{x})$, that captures the expected value of f at each point in $\mathbf{x}$, and a kernel function, $k(\mathbf{x}, \mathbf{x}')$, that specifies the covariance between outputs $f(\mathbf{x})$ and $f(\mathbf{x}')$ as a function of inputs $\mathbf{x}$ and $\mathbf{x}'$. These two objects determine the distribution of f, such that, for a fixed set of inputs $\mathbf{x}_1, \ldots, \mathbf{x}_N$: $f(\mathbf{x}) \sim \mathcal{N}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$. The properties of the functions are governed by the assumed mean and kernel functions. For example, if the kernel function were decreasing in the distance between $\mathbf{x}$ and $\mathbf{x}'$, then function values corresponding to nearby inputs will be more correlated than those corresponding to more distant inputs. This property is useful in time series and other dynamic models where the influence of an early observation on a later one decays over time (Rasmussen and Williams 2006), and more generally in cases where we expect the function of interest to be relatively smooth. The rate of this distance decay may be determined by hyperparameters within the kernel, thus determining the overall smoothness of the function.[9]

While GPs have been common in the broader PML community for decades, they have only recently seen use in marketing. Dew and Ansari (2018) use GPs to enrich models for customer base analysis, allowing for the presence of potentially unknown calendar time disruptions in repurchase rates, by modeling the repurchase likelihood as an unknown function of four customer-level "time scales," including calendar time. In a similar vein, Dew et al. (2024) develop a novel GP-based method to isolate customer-level routines from transaction data, using a novel kernel to embed prior knowledge about how transactions may be related across hours and days. Beyond predicting purchasing, GPs have also been used to relax assumptions in latent utility models. For instance, Dew et al. (2020) develop a GP-based specification of *dynamic heterogeneity* that allows for individual-level preference parameters to evolve over time, by modeling them with hierarchical GP priors: $\boldsymbol{\theta}_i(t) \sim \mathcal{GP}(\mu(t), k(t, t'))$. This hierarchical specification regularizes the function values, allowing for individual-level inference even with limited observations. Other work has built

[8] More technically, a GP is a scalar-valued stochastic process $f(\cdot)$ over an input space, $\mathcal{X}$, such that, for any set of inputs, the corresponding values $f(\mathbf{x})$ are a realization of a multivariate Gaussian random variable.

[9] Infinitely wide deep neural networks are equivalent to GPs (Lee et al. 2017), highlighting their immense flexibility.

on the idea of hierarchical GP methods for capturing utility: Dew (2023) uses individual-level GPs in preference measurement to model customer-level utility over unstructured data, and Korganbekova and Zuber (2023) uses GPs to model unknown utility functions in models of consumer search. In other applications, GPs have been used to relax functional form specifications in direct utility models (Levy and Montgomery 2024), models that impute credit risk (Tian 2019), and models of preferences for charitable giving (Kim et al. 2021). Across these contexts, GPs are used for their regularized flexibility, which can capture complex patterns in how inputs drive utility, and the ease of incorporating GPs in existing models.

Bayesian Splines An alternative approach to estimating unknown functions is through Bayesian variants of splines. Splines are common in generalized additive models, wherein an outcome Y_i is modeled as an additive combination of unknown functions: $Y_i = \sum_{j=1}^{J} f_j(X_{ij}) + \varepsilon_i$. Bayesian variants of splines include, for example, piecewise second-order polynomials, given by:

$$f_j(X_{ij}) = \beta_{j0} + \beta_{j1}X_{ij} + \beta_{j2}X_{ij}^2 + \sum_{q=1}^{Q} \beta_{jq}(X_{ij} - \kappa_{jq})_+^2,$$

where $\beta_{j0}, \beta_{j1}, \beta_{j2}$ and $(\beta_{jq})_{q=1}^{Q}$ are parameters and $\kappa_{j1} < \cdots < \kappa_{jQ}$ are fixed knots. A key benefit of modeling splines probabilistically is the ability to regularize all components of the model through informative priors. For instance, the regularization of $(\beta_{jq})_{q=1}^{Q}$ ensures that only important knot parameters have significant effects, echoing the connections between feature discovery methods like LASSO and Bayesian regression discussed in Section 2.2. This is the spline specification used in Boughanmi and Ansari (2021), who model the effects of acoustic features on the success of a music piece. In other marketing contexts, Kalyanam and Shively (1998) use stochastic spline regressions within a hierarchical Bayes model to estimate irregular pricing effects. Kim et al. (2007) use splines to develop flexible, individual-level utility functions to model choice, while accommodating individual-level monotonicity in price response.

Bayesian Neural Networks Given the immense success of neural networks in estimating unknown functions, an emerging literature in PML explores the possibility of *Bayesian* neural networks (BNNs), which merge neural network architectures and Bayesian inference (see Papa-

markou et al. 2024 for a review). In a BNN, both weights and biases within the neural network architecture are modeled as random variables. Contrary to traditional neural networks, which commonly only offer point estimates, BNNs capture the uncertainty inherent in both the parameters and the predictions, thus making them well-suited for tasks requiring decision-making under uncertainty, with less extreme data requirements. For instance, Daviet (2020) shows the potential utility of BNNs for modeling the effect of product images on consumer preferences, in a case where there are relatively few images. In marketing, the applications of BNNs are so far limited, and offer a fertile area for future research.

3.3 High-Dimensional Data

One of the hallmarks of modern data is high dimensionality: in addition to "long" data involving observations of many individuals, marketers often deal with "wide" data where many different variables are observed for the same individual or the variables themselves are complex or unstructured. PML models have seen great success in helping marketers utilize these types of data, extracting meaningful low-dimensional structure from these large and complex sets of variables and allowing for statistically efficient predictions of future outcomes. We overview three classes of PML models for such problems: matrix factorization models of dyadic interactions, embedding models of co-occurrences, and deep generative models of unstructured data.

Matrix Factorization for Dyadic Outcomes Many marketing applications involve modeling dyadic outcomes among a potentially large collection of units (e.g., customers, products). For example, we may have data on how many times X_{ij} that customer i has purchased product j for a large set of products, $j = 1, 2, \ldots, J$, over a fixed period of time on an e-commerce platform, and want to predict future purchases, particularly for unseen customer-product pairs, to make recommendations. Such datasets are often high-dimensional and sparse: on an e-commerce platform with thousands of products, customers will generally have only purchased a tiny fraction of the available products. While hierarchical models of unobserved heterogeneity like those discussed in Section 3.1 are suitable for estimating individual-level preferences with limited observations, they are impractical in such high-dimensional settings. Instead, we can model such data using latent spaces, and in particular, matrix factorization models.

Latent space models represent each customer and product as a point (or "embedding") in a lower-dimensional latent space, and model outcomes as a function of those embeddings. Intuitively, units that tend to have similar outcomes (e.g., consumers who tend to buy similar products) are estimated to be close together in the latent space, and thus will be predicted to have similar outcomes in unseen dyads, allowing information pooling both across customers and across products. The dimension of the latent space is typically chosen to be much lower than the number of customers or products, resulting in a drastic reduction in parameters to be estimated. The inferred latent space tends to have meaningful structure, with distances capturing intuitive notions of similarity and directions in the space capturing attributes, aiding in interpretability. Latent space models in bipartite settings like this, where interactions are modeled between two distinct collections (i.e., customers and products), are commonly referred to as matrix factorization models, since the data can be represented in a (potentially partially missing) $N \times J$ matrix of observations.[10] The embeddings can be seen as factorizing that matrix into a product of low-rank matrices, analogous to matrix decompositions like the singular value decomposition (Koren et al. 2009). Though originally conceived based on linear algebraic and geometric intuitions, in PML, the embeddings are treated probabilistically as latent variables, which has naturally led to extensions tailored towards data with specific distributional properties (e.g., sparsity, discreteness, and overdispersion) and more efficient inference algorithms.

For example, a useful form of PML matrix factorization for modeling customers and products is *Poisson factorization* (PF; Canny 2004; Gopalan et al. 2015). In PF, we represent each customer and product by K-dimensional, non-negative embedding vectors $\boldsymbol{\theta}_i$, $\boldsymbol{\phi}_j \in \mathbb{R}_+^K$, and model the purchase count X_{ij} as Poisson-distributed with the rate parameter equal to the dot product of the embedding vectors: $X_{ij} \sim \text{Poisson}\left(\boldsymbol{\theta}_i^\top \boldsymbol{\phi}_j\right)$. In turn, the embedding vectors $\boldsymbol{\theta}_i$, $\boldsymbol{\phi}_j$ are modeled as being drawn coordinate-wise from independent Gamma priors, which induces sparsity in the embeddings. $\boldsymbol{\phi}_j$ can be seen as capturing latent attributes of each product, and $\boldsymbol{\theta}_i$ can be seen as capturing consumer preferences in terms of which attributes they tend to purchase. These

[10] Beyond the bipartite setting, embeddings of units into lower-dimensional latent spaces have also seen popularity in modeling interactions within a single collection of units. For instance, analyses of social networks have modeled individuals as occupying a location $\boldsymbol{\theta}_i$ in a latent "social space," with dyads of individuals closer together in this space being more likely to be connected to each other, capturing the principle of homophily (Hoff et al. 2002; Braun and Bonfrer 2011). The latent space is often intuitively interpretable, showing how individuals group into social clusters or cliques and the relationships among them.

latent variables are learned entirely from purchase outcomes based on which products tend to be purchased by the same people, and thus do not require any product-specific attribute data.[11] Poisson factorization has also been used as a topic model for text, with X_{ij} giving the number of times word j is used in document i.

In marketing, Poisson factorization and extensions thereof have been used to model purchase behavior (Liu and Kawaguchi 2022) as well as to model topics in textual data (Toubia 2021; Liu et al. 2021). One limitation of the basic matrix factorization approach is that, while it can yield insights about consumer preferences in terms of purchase outcomes and can predict future purchase outcomes, the inferred user latent variables are not directly interpretable as preferences in the sense of microeconomic utility (as a multinomial logit is), and the model may not be able to predict counterfactual outcomes under different scenarios (e.g., different prices or product assortments). To this end, recent work in economics and marketing combine matrix factorization approaches with random utility models, explicitly modeling price sensitivity and product complementarity with a latent embedding structure (Ruiz et al. 2020; Donnelly et al. 2021). These models retain the microeconomic interpretation of choice models and the ability to predict counterfactuals, while allowing the model to scale to massive datasets with thousands of products.

Embedding Models of Co-Occurrences Another common application of latent space models is in modeling assortments of items that co-occur together, such as products in a shopping basket, words in a sentence, or ingredients in a recipe. For instance, suppose an online grocery shopping platform has data on many shopping baskets purchased by past customers. The platform wants to understand the role that each product plays in a basket assortment, such as which products are substitutes and complements, e.g., to recommend replacements for out-of-stock products and to recommend complements to complete a basket. Likewise, modeling the co-occurrences of words within a sentence can help researchers understand the semantic role that each word plays in a sentence and which words are likely to co-occur with each other.

Similar to matrix factorization models, we can approach this problem by embedding each unit (e.g., a product or a word) into a latent space, modeling co-occurrences between pairs of units as

[11] One particularly appealing aspect of PF is computation: Gamma-Poisson conjugacy allows for closed form complete conditionals for the latent parameters that depend only on non-zero observations, which in turn, allows for efficient implementation of posterior inference.

a function of their embeddings. Embedding models of co-occurrences have seen great success, to the point that they are commonly simply referred to as "embedding models." The most successful of these models is the word2vec model for learning embeddings of words (Mikolov et al. 2013), which models the likelihood of a word appearing in a given context (i.e., neighboring words) by means of low-dimensional embeddings. Word embeddings have enjoyed tremendous popularity for their ability to capture semantic information about the meaning and usage of words in a relatively low-dimensional, continuous space. While many embedding models are not strictly probabilistic (i.e., do not have a proper DGP), they are often important ingredients in model building and data processing (e.g., Timoshenko and Hauser 2019).

PML has contributed a number of specifications beyond word embeddings, that can be specifically tailored to modeling products. Most notable are the exponential family embeddings of Rudolph et al. (2016), which generalize word embeddings to allow the context and target variables to come from general exponential family distributions. For example, Poisson-distributed variables can be used to model the quantity of an item purchased in a shopping trip conditional on the other items in the basket. In marketing, Sozuer et al. (2024) use exponential family embeddings to model the roles of ingredients in recipes, using the learned embeddings to characterize the creativity of recipes and relate the creativity of a recipe to its adoption and quality. Other types of product embeddings have been developed by Ruiz et al. (2020), who use product embeddings in their choice utilities to capture how previously added products in a basket help predict the next product to be added, and by Chen et al. (2022), who introduce a product embedding method to study product-level competition.[12]

Deep Generative Models of Unstructured Data In addition to data with many variables per observation (e.g., a large collection of products), another form of high-dimensional data is so-called "unstructured" data, where the variables themselves are complex objects rather than simple scalar or categorical variables. Unstructured data like images (Feng et al. 2023), audio (Fong et al.

[12]One conceptual difficulty with these co-occurrence based models is that they are not proper probabilistic generative models: the likelihood of a given target is specified conditional on its context, but because the contexts can in general be non-nested across observations, the terms cannot be combined into a coherent joint likelihood (Rudolph et al. 2016). Thus, these models cannot simulate new data, the posteriors of the embedding and context vectors are not well-defined. This is particularly problematic for unordered assortments like consumer basket choice. Solutions include heuristic approximations (Ruiz et al. 2020) or exact but computationally expensive data augmentation (Kosyakova et al. 2020).

2023), and video (Yang et al. 2023) are crucial for firms (e.g., to design advertisements/products or to manage brand image/online presence) and for consumers (e.g., to express emotions, thoughts, and preferences on social media/review platforms). Given their importance, how can one model and extract insights from such data? From a probabilistic perspective, analyzing unstructured data centers once again around the DGP. Intuitively, by building models that can generate unstructured data, we can use the latent variables of those models as meaningful representations, which may in turn be used for downstream tasks, or integrated with other models.

The challenge of modeling unstructured data is dimensionality: textual data features a multitude of unique words, while image data comprises immense grids of pixel values.[13] To reduce this dimensionality, neural network architectures have been designed to process unstructured data, like convolutional neural networks for images or transformer models for text. These tools have primarily been used for prediction, though they may also be used as components of generative models, granting the ability to produce and manipulate new unstructured data.[14] To actually build a generative model—that is, a model $p(\mathcal{D}_i \mid \boldsymbol{\theta}_i)$, where $\mathcal{D}_i$ is an observation of unstructured data, and $\boldsymbol{\theta}_i$ is a latent variable capturing the underlying features of that observation—the marketing literature has largely turned to two PML specifications: variational autoencoders (VAEs) and generative adversarial networks (GANs).

VAEs (Kingma and Welling 2013; Rezende et al. 2014) model an observation $\mathcal{D}_i$ as generated from a latent variable $\boldsymbol{\theta}_i$, with conditional density $p_{\boldsymbol{\omega}}(\mathcal{D}_i \mid \boldsymbol{\theta}_i)$ specified as a neural network parameterized by $\boldsymbol{\omega}$ (termed the decoder or generative network). While this parameterization is highly flexible, it makes posterior inference of $p(\boldsymbol{\theta}_i \mid \mathcal{D}_i)$ difficult, due to the complex structure of the generative model. This difficulty is addressed by approximating the posterior with a simpler distribution $q_{\boldsymbol{\phi}}(\boldsymbol{\theta}_i \mid \mathcal{D}_i) \approx p(\boldsymbol{\theta}_i \mid \mathcal{D}_i)$, which itself is specified using another neural network parameterized by $\boldsymbol{\phi}$ (termed the encoder/inference network). Thus, VAEs consist of two neural networks that play inverse roles to each other: $q_{\boldsymbol{\phi}}$ maps observations $\mathcal{D}_i$ to a conditional distribution over latent

[13] More specifically, at the data level, an image is an array of pixel values. For color images, there are typically three channels per pixel, capturing the intensity of various colors. Thus, even small images are big data: a tiny 100 x 100 pixel color image, which at standard resolutions, would be a half-inch square, contains 30,000 values.

[14] Although large language models (LLMs) can generate text, they rely on ad hoc techniques like temperature adjustment and nucleus sampling to produce more random, diverse content and have imperfectly calibrated uncertainty (Xiao et al. 2022). To this end, LLMs could benefit from a probabilistic perspective, which may be helpful managerially, for example, in delivering more accurate interventions that use confidence-based text highlighting to alert consumers to potential hallucinations (Spatharioti et al. 2023).

variables $\boldsymbol{\theta}_i$, while $p_{\boldsymbol{\omega}}$ maps latent variables to a conditional distribution over observations.[15] The architectures of both neural networks can take any form, including deep convolutional neural networks, allowing VAEs to learn to represent and generate complex data like images.[16] Moreover, the parameters $\boldsymbol{\phi}$ are "amortized" or shared across all observations, permitting considerable scalability, which is crucial when working with large volumes of high-dimensional data (see Section 4.1 for details).

VAEs have been leveraged for several marketing applications involving unstructured data. Dew et al. (2022) extract a meaningful set of image features describing logos, and then use a multimodal VAE to understand how those features connect to textual aspects of the brand and consumer perceptions. Tian et al. (2023b) follow a similar procedure to learn representations of content in influencer marketing. Boughanmi et al. (2023) use VAEs to analyze consumer collections like music playlists and demonstrate how their model can generate samples of novel playlists. Cheng et al. (2022) leverage VAEs to create representations from patent text, which are then used to construct economically interpretable measures. Sisodia et al. (2023) apply VAEs to extract interpretable representations directly from product images and characterize preferences over their latent attributes using conjoint analyses. Here, generative modeling is particularly helpful, since new product designs can be created to target specific consumer segments.

An alternative way to probabilistically generate unstructured data is through GANs. Since VAEs define and estimate $p(\mathcal{D}_i \mid \boldsymbol{\theta}_i)$, they are considered models of explicit density estimation. In contrast, GANs (Goodfellow et al. 2014) address implicit density estimation, sampling from $p(\mathcal{D}_i \mid \boldsymbol{\theta}_i)$ without directly solving for it. GANs are defined by two neural networks: a generator G and a discriminator D, parameterized by $\boldsymbol{\phi}$ and $\boldsymbol{\omega}$, respectively. The generator produces a data point from $\boldsymbol{\theta}$, which is sampled from a standard normal prior $p(\boldsymbol{\theta})$. The discriminator then determines the probability that this data point is real (i.e., from the observed data $\mathcal{D}$) rather than fake (i.e., from the generator). The model's objective function is a minimax (or adversarial) game,

[15] This motivates the name "VAE," since the model structure is a probabilistic analog of a traditional autoencoder (Rumelhart et al. 1986), which consist of an "encoder" network that maps observations into a lower-dimensional latent vector and then a "decoder" network that reconstructs the original observation from the latent vector. Compared to traditional autoencoders, VAEs tend to learn latent representations that are more continuous and well-structured (due to regularization by a smooth prior on $\boldsymbol{\theta}_i$, typically a standard Gaussian distribution), are capable of generating new data, and founded on clear statistical grounds.

[16] Notably, diffusion models, popularized by text-to-image generative models like DALL-E, Stable Diffusion, and Midjourney, can be characterized as a type of hierarchical VAE (Luo 2022).

in which the generator tries to fool the discriminator by generating data indistinguishable from $\mathcal{D}$:

$$\min_{\phi} \max_{\omega} \mathop{\mathbb{E}}_{\mathcal{D}\sim p_{\text{data}}} \Big[\log \underbrace{D_{\omega}(\mathcal{D})}_{\substack{\text{recognize real} \\ \text{data as "real"}}} \Big] + \mathop{\mathbb{E}}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta})} \Big[\log \Big(\underbrace{1 - D_{\omega}\big(\overbrace{G_{\phi}(\boldsymbol{\theta})}^{\substack{\text{generate fake data} \\ \text{that look "real" to } D}} \big)}_{\substack{\text{recognize fake} \\ \text{data as "fake"}}} \Big) \Big] \tag{7}$$

In marketing, GANs have seen limited but growing adoption. Burnap et al. (2023) use a VAE-GANs hybrid model trained directly on product images to predict aesthetic scores and generate novel appealing designs. Anand and Lee (2023) use GANs to generate artificial customer data, set price markups, and target customers. Luo and Toubia (2024) use GANs to generate realistic faces and then manipulate their femininity to address questions about gender discrimination based on facial features.

3.4 Missingness and Data Fusion

Marketing applications often need to leverage information from multiple disparate data sources simultaneously, with observations either being unlinked or imperfectly linked across data sources (e.g., due to different sample selection or lacking a common identifier across datasets). For instance, suppose a firm is interested in understanding the relationship between media viewing and product purchasing to determine which products should be marketed on which channels (Gilula et al. 2006). Though the firm may have data on the purchasing behavior of their own customers, they will need to rely on an external data provider for media viewing (e.g., a business intelligence firm that conducts consumer surveys), and it will generally not be possible to link individuals in the external data to those in the internal data. This can be further complicated by data being at differing levels of aggregation (e.g., aggregate media viewership data; Feit et al. 2013), as well as potential selection bias resulting in one or more data sources being non-representative of the target population of interest (McCarthy and Oblander 2021; De Bruyn and Otter 2022).

In marketing, data fusion refers to jointly modeling multiple data sources to make inferences that are more accurate, generalizable, and useful than could be achieved with any single dataset alone. This is typically done by modeling all data sources conditional on a set of common variables and estimating the model on all data sources jointly. For data fusion with unlinked data,

meaning there is no common identifier across the datasets, prior work has achieved identification by assuming conditional independence of the outcomes of interest across data sources (e.g., media viewership and product purchases), conditional on common auxiliary variables observed in both datasets, such as demographics (Gilula et al. 2006; Feit et al. 2010; Qian and Xie 2014). The PML perspective has the potential to offer more flexibility by linking variables of interest via common *latent* variable structures. For example, suppose $\mathbf{X}_i$ consists of one consumer's choices in a conjoint study, while $\mathbf{Y}_{i'}$ consists of another consumer's purchasing decisions in a panel. When there is overlap in the variables observed or common structure can be assumed in the data generating process (e.g., stable preferences across conjoint and scanner panel choice data), the two can be mapped together, by assuming both sets of variables come from the same underlying latent structure: $p(\mathbf{X}_i \mid \boldsymbol{\theta}_i)$, $p(\mathbf{Y}_i \mid \boldsymbol{\theta}_{i'})$. This allows us to estimate a joint model of both types of variables, mapping them to the same latent consumer preferences, $\boldsymbol{\theta}_i$. Aggregated data can be handled by directly approximating the likelihood of the aggregate data (McCarthy and Oblander 2021) or using Bayesian imputation (Feit et al. 2013). If there are some linked observations (i.e., some consumers for whom both conjoint and scanner panel choices are observed), these joint observations enable identification, ensuring the inferred $\boldsymbol{\theta}_i$'s are comparable across the two types of data.

In marketing, PML approaches to data fusion have been applied in a number of contexts. A particularly powerful application of PML data fusion is for utilizing multiple types of unstructured data. Multiview generative models treat multiple different modes of observations as coming from the same latent space, allowing for translation between modes (e.g., brand logos, textual brand descriptions, and consumer perceptions of brand personality; Dew et al. 2022). More generally, PML is well-suited to jointly modeling multiple processes or behaviors, which can provide more precise estimates of a customer's preferences even when there are limited observations of any one behavior, such as in customer relationship management (Padilla and Ascarza 2021) and web search (Liu et al. 2021) contexts. PML methods have also been successful at handling partial missingness of some variables, as is common in consumer data. In particular, if variables are not missing completely at random (e.g., survey respondents do not respond to questions they are indifferent about), then missingness of one variable can provide information about other variables. Probabilistic latent variable models can allow for selection on unobservables by modeling selection jointly with the main variables of interest using common latent variables (Tian and Feinberg

2020), resulting in posterior inferences of the latent variables and their population distributions that correct for selection bias. Finally, and most recently, probabilistic data fusion methods have shown great promise for protecting customer privacy when fusing potentially sensitive data (Tian et al. 2023a).

4 How To Do It: Computation and Implementation

As described in Sections 1 and 2, the heart of PML is Bayesian inference, which provides a comprehensive picture of uncertainty through posterior distributions, rather than just point estimates. This distinction is crucial in marketing where understanding the range of possible outcomes is often as important as predicting the single most likely scenario. However, computation is a challenge: as noted in Section 2.3, while Bayes' theorem provides a natural mechanism for inference in theory, in practice the posterior is almost always impossible to compute. Thus, in this section, we show how to combine models from Section 3 with computational tools to compute the posterior in practice. We highlight the important advances in Bayesian computation that have not only enabled the development of PML, but also made it easier to adopt. Finally, to showcase the power and ease-of-use of these tools, we also include a code companion demonstrating them context of the mixed logit model from Section 1, which can be accessed at: `https://rtdew1.github.io/code_appendix.html`.

4.1 Overview of Methods

There are two broad categories of popular inference methods: sampling-based methods and approximation methods. Sampling-based methods, specifically MCMC methods, aim to generate samples from the posterior distribution of the parameters of interest, despite the posterior density only being known up to proportionality. The idea of MCMC is to create a Markov chain based on the joint density of data and parameters, such that the distribution of the samples from the chain converges to the target posterior distribution. Until very recently, the vast majority of Bayesian models in marketing relied on MCMC for inference, and specifically on two such algorithms: random walk Metropolis-Hastings (RWMH) and Gibbs Sampling (GS). Though these algorithms still play an important role in computation, they are limited: RWMH is a very general algorithm,

requiring no model-specific derivations, but convergence to the posterior can be slow, and computation can be prohibitive for large models. On the other hand, GS is often more efficient, but involves model-specific derivations and is only compatible with a limited class of conditionally conjugate models, which excludes most of the models discussed in Section 3.

The advent of modern computing, especially fast algorithms for differentiation and optimization, has led to a host of new algorithms that alleviate the limitations of RWMH and GS, and enabled the rise of PML. Here, we will review three of the most important innovations: (1) gradient-based sampling methods, (2) stochastic variational inference, and (3) amortized variational inference. We also summarize these methods in Table 1.

Hamiltonian Monte Carlo Hamiltonian Monte Carlo (HMC) is an MCMC method that leverages the gradient of the log joint density of the model's data and parameters (i.e., $\log p(y, \theta) = \log p(y \mid \theta) + \log p(\theta)$), to design a more efficient Markov chain for sampling from the posterior (Neal 2011). Inspired by the dynamics of physical systems, HMC combines this gradient with auxiliary momentum parameters that guide the steps of the sampler, leading it to quickly converge to regions of high posterior mass. Rather than a fully random walk, HMC simulates the movement of a particle around the parameter space, where the negative log-joint represents the potential energy state of the particle. Intuitively, at each HMC step, the particle is "kicked" in a random direction and the movement of the particle is governed by a combination of the momuntem from this kick and the shape of the log-joint function according to classical (Hamiltonian) dynamics. This Markov chain tends to mix much faster than RWMH, while retaining the posterior as the stationary distribution. Crucially, HMC can be implemented using automatic differentiation tools, which have become widely available within ML toolkits, thus requiring no model-specific derivations. Moreover, while standard HMC requires carefully tuning the values of several hyperparameters, modern variants of HMC—most notably, the No-U-Turn Sampler (or NUTS)—use heuristics to find good settings as part of the training process (Hoffman and Gelman 2014). Thus, researchers can perform HMC-based inference by simply specifying the joint log-joint of the data and model parameters. This ease of use has led to the wide adoption of HMC in many Bayesian inference libraries, especially the probabilistic programming languages (PPLs) we will discuss later.

While HMC has seen sporadic use in marketing in the past (e.g., Qian and Xie 2011), the ease

of using HMC through PPLs has led to an explosion of marketing papers adopting it in recent years (e.g., Dew and Ansari 2018; Tian and Feinberg 2020; Padilla and Ascarza 2021; Karlinsky-Shichor and Netzer 2023). While powerful, HMC is also limited: because it relies on gradients of the log-joint, it can only be used to infer the posterior in models with continuous parameter spaces. Moreover, while efficient, HMC still relies on repeatedly evaluating the likelihood for all observations and jointly sampling the entire parameter vector, which can be computationally burdensome with massive data. While there have been some attempts to alleviate this burden for massive data (e.g., through stochastic gradients), alternative methods, like those discussed subsequently, may be more appropriate for such cases.

Variational Inference In contrast to the sampling methods described previously, Variational Inference (VI) transforms the problem of Bayesian inference into an optimization problem, by approximating the true posterior $p(\boldsymbol{\theta} \mid \mathcal{D})$ with a simpler distribution $q(\boldsymbol{\theta})$, and optimizing q to be as "close" to the posterior as possible (Blei et al. 2017). In VI, we specify a family of approximating distributions $\mathcal{Q}$ (i.e., the variational family) and then find the distribution q^* within this family that minimizes the Kullback-Leibler divergence (KLD) from the posterior:

$$q^*(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta}) \in \mathcal{Q}}{\arg\min}\, KL\big(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathcal{D})\big) = \underset{q(\boldsymbol{\theta}) \in \mathcal{Q}}{\arg\min}\, \mathbb{E}_q \left[\frac{\log q(\boldsymbol{\theta})}{\log p(\boldsymbol{\theta} \mid \mathcal{D})} \right]$$

We cannot compute the KLD in practice, since it depends on the (intractable) posterior $p(\boldsymbol{\theta} \mid \mathcal{D})$. However, the KLD is equivalent to (the negative of) $\mathbb{E}_q\big[\log p(\mathcal{D} \mid \boldsymbol{\theta})\big] - KL\big(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta})\big)$, up to an additive constant. This expression is referred to as the "evidence lower bound" (ELBO), as it lower bounds the evidence $p(\mathcal{D})$, with the bound being tight when $p(\boldsymbol{\theta}_i \mid \mathcal{D}) \in \mathcal{Q}$ (Braun and McAuliffe 2010). Crucially, the ELBO depends only on known densities, and thus can be computed (or at least approximated) efficiently. Thus, variational inference finds the distribution q^* that maximizes the ELBO. Most commonly, the variational family $\mathcal{Q}$ is restricted to *mean field* families, where the dimensions of the distribution q (i.e., the individual latent variables) are assumed to be statistically independent, which greatly simplifies optimization.

In conditionally conjugate cases, it is often possible to derive the optimal mean field variational approximation for one variable conditional on the variational approximations of all other

variables, allowing for efficient coordinate ascent VI (CAVI), which cycles through updating the optimization variational distribution for each variable. Much like Gibbs sampling, this can be efficient, but is only feasible for conditionally conjugate models and requires model-specific derivations. As such, modern VI methods focus on "black-box" optimization, where the variational family and model likelihood are both allowed to be generic distributions without requiring conjugacy or model-specific derivations. This is where the advantages of framing the inference problem as an optimization problem shine: because the inference problem is posed as simply minimizing an objective function, many of the optimization tools that have enabled modern ML frameworks to scale can be directly applied to VI. Most notable is the use of stochastic gradients, where a noisy approximation to the gradient is used instead of the exact gradient in a gradient-based optimizer.[17] This combination is referred to as Stochastic Variational Inference (SVI), which has successfully enabled inference for complex Bayesian models across a variety of contexts.

The literature on VI is vast, and recent innovations in automatic differentiation have been paired with variance control methods to develop automatic variational inference algorithms, including black-box variational inference (BBVI), automatic differentiation variational inference (ADVI), and pathfinder variational inference (PVI), which all allow for fast estimation of generic models without requiring model-specific derivations (Ranganath et al. 2014; Kingma and Welling 2013; Kucukelbir et al. 2017; Zhang et al. 2022).[18] Though VI is scalable and flexible, it also does not have the same theoretical guarantees as sampling-based methods like MCMC. In particular, the KLD objective function incentivizes mode-seeking behavior, wherein q^* tends to place disproportionate probability mass on high posterior density points, resulting in understatements of model uncertainty. This is especially pronounced when the variational family is misspecified. Theoretically, Wang and Blei (2018) show that the variational posterior's mean asymptotically converges to the true value of the latent variable, but that its variance asymptotically underestimates the true posterior variance.[19] As such, other work has sought to improve the quality of the approximation

[17] As commonly used in ML, stochastic gradients typically refer to subsampling a random set of observations and computing the gradient for this random subset. In variational inference, there is often a second layer of stochasticity: the expectations over q in the ELBO often do not have a closed form solution, so random simulations from the variational distribution are used to approximate them.

[18] A particularly successful variance control method is the "reparameterization trick" of Kingma and Welling (2013).

[19] In the authors' experience, this underestimation of variance is more severe for global, population-level parameters, like a mean, and less problematic for local, individual-level parameters, like a customer's price sensitivity. As many marketing decisions rely on the latter, such systematic biases might not dramatically affect decisions. We illustrate this through simulation in the accompanying code companion.

by allowing for more complex variational families that can be non-normal and correlated across dimensions. Particularly popular are normalizing flows (Papamakarios et al. 2021), wherein a simple base distribution is transformed through a series of invertible transformations to a more complex one. The parameters of these transformations are then optimized as part of (S)VI, enabling more accurate variational approximations (Rezende and Mohamed 2015).

VI methods have seen moderate adoption in marketing, especially in recent years. Braun and McAuliffe (2010) derive a variational family and efficient VI estimation procedure to allow hierarchical multinomial logit choice models to scale to a large number of individuals. Dzyabura and Hauser (2011) use VI to adaptively select conjoint questions to infer what heuristic decision rules consumers are using. More recently, other researchers have developed novel CAVI algorithms to efficiently estimate more complex probabilistic models: e.g., Ansari et al. (2018), to build hybrid recommender systems; Xia et al. (2019), to model complex shopping patterns; Toubia (2021), to study creative documents; Liu et al. (2021), to model web search behavior; Jacobs et al. (2021), to understand large-scale purchase behavior; and Donnelly et al. (2021), to estimate preferences based on high-dimensional discrete choices.

Amortized Inference A third noteworthy innovation in the space of Bayesian computation is amortized (variational) inference for local variables. The idea of amortized inference is, instead of separately optimizing the variational distribution $q(\boldsymbol{\theta}_i)$ for every individual i, to "amortize" the computational cost of optimization across units by estimating a single function $q_\phi(\boldsymbol{\theta}_i; \mathbf{X}_i)$ that maps observations $\mathbf{X}_i$ to a variational posterior over $\boldsymbol{\theta}_i$, parameterized by ϕ. The parameters ϕ are chosen to maximize the ELBO, and all the same computational techniques used for standard VI (e.g., stochastic optimization) are applicable (Kingma and Welling 2013; Rezende et al. 2014). $q_\phi(\boldsymbol{\theta}_i; \mathbf{X}_i)$ is often parameterized by a neural network, commonly referred to as the "inference network," or as the "encoder" in the context of VAEs, as described in Section 3.3.

Amortized inference results in much more scalable models for data with large numbers of local parameters to be estimated, since it allows for optimization of a single function mapping datapoints to posteriors, as opposed to solving a large number of independent optimization problems for every observation. This also makes it easy to perform out-of-sample inference of posterior parameters for new datapoints, since new observations can simply be plugged into the estimated

function. The downside of amortization is that the quality of the variational approximation now depends not only on the quality of the variational family in approximating the true posterior, but also the quality of the inference network in capturing the relationship between data and posterior parameters. Amortized variational inference has been leveraged in a number of marketing applications of VAEs (see Section 3.3). In particular, Dew et al. (2022) modify the way inference works in a multimodal VAE to allow for different patterns of missingness in the inputs, enabling decision support tools for designers and managers; and Boughanmi et al. (2023) use a hypergraph convolutional neural network to perform posterior inference on the embeddings of multiple levels of collections of individual units (e.g., songs within a playlist or a collection of playlists made by a listener).

Method	Advantages	Disadvantages	Citations
Random Walk Metropolis-Hastings (RWMH)	- Can be applied to virtually any model - Requires no model-specific derivations	- Slow convergence to the posterior - Computationally prohibitive for large models	Hastings (1970) e.g., Allenby et al. (1998)
Gibbs Sampling (GS)	- Very efficient when applicable	- Only applicable to specific models with specific "conjugate" priors - Requires model-specific derivations	Geman and Geman (1984) e.g., Rossi et al. (1996)
Hamiltonian Monte Carlo (HMC)	- Leverages gradient of model's log-density for efficient sampling - Quick convergence to high posterior mass regions - When paired with automatic differentiation tools, no model-specific derivations needed	- Limited to continuous parameter spaces - Slow and/or computationally burdensome with massive data	Neal (2011); Hoffman and Gelman (2014) e.g., Qian and Xie (2011); Dew and Ansari (2018)
Variational Inference (VI)	- May be faster than sampling-based methods, especially with stochastic gradients (SVI) - Compatible with modern machine learning frameworks	- Requires careful choice of variational family - Approximation quality depends on this choice	Jordan et al. (1999); Blei et al. (2017) e.g., Braun and McAuliffe (2010); Ansari et al. (2018)
Amortized Inference	- Very efficient for large datasets - Suitable for complex models with neural network implementations	- Quality of inference depends on the accuracy of the learned function *and* the variational family	Kingma and Welling (2013); Rezende et al. (2014) e.g., Dew et al. (2022)

Table 1: Comparison of Inference Methods

4.2 Probabilistic Programming Languages

One of the most important developments in the Bayesian computation literature is the rise of probabilistic programming languages, or PPLs. PPLs allow for the specification of probabilistic mod-

els in simple terms, often through a function written in a standard programming language (e.g., Python) specifying the model's (log) density function (i.e., $p(\mathbf{x}, \boldsymbol{\theta})$). At a high level, PPLs work by using automatic differentiation to compute the gradient of the model's log-joint, then using efficient implementations of either NUTS or black-box variational inference to perform Bayesian inference on the model. These frameworks have dramatically simplified the Bayesian inference pipeline: now, rather than needing to develop both the model and a corresponding inference algorithm, researchers need only be able to write a function that computes the model's log-joint, which is typically straightforward. This ease facilitates not only the development of a single model, but the testing of various specifications, as the code for the log-joint can be modified without needing to make extensive changes to the inference procedure. In many cases, PPLs are based on deep learning libraries like PyTorch and Tensorflow, which enables the development of PML models that fuse deep learning ingredients (e.g., automatic differentiation, stochastic optimization, and efficient matrix operations) with probabilistic modeling and Bayesian inference. We summarize state-of-the-art PPLs in Table 2. For most researchers and practitioners, PPLs based on NUTS (e.g., Stan, PyMC) are an accessible starting point, while PPLs based on variational inference and more complex MCMC algorithms (e.g., Pyro, Tensorflow Probability) present opportunities for improving scalability. We demonstrate both Stan and Pyro in the web appendix.

Language	Interface	Algorithms	Backend	Comments
Stan	Custom language (callable from R, Python, Julia, MATLAB, and Stata)	HMC, NUTS, ADVI, PVI	Custom automatic differentiation library, built on C	Most beginner-friendly; includes easy-to-use extensions for specific models, like rstanarm for applied regression
PyMC	Python	MH, HMC, NUTS, ADVI, Stein VI, Sequential Monte Carlo (SMC)	Aesara and PyTensor	Supports combining samplers, so that MH can be used to sample discrete variables; beginner-friendly.
Pyro	Python	BBVI, HMC, NUTS, Stein VI, SMC, Reweighted Wake-Sleep	PyTorch	Includes automatic implementations of many variational families, including normalizing flows. Highly customizable, especially with BBVI. Allows for amortization. Challenging for beginners.
NumPyro	Python	HMC, NUTS, Mixed HMC, BBVI	JAX	Originally built as a JAX-based NUTS-specific version of Pyro. Mixed HMC allows for the inclusion of discrete variables. Allows for amortization. Challenging for beginners.
TensorFlow Probability	Python	MH, HMC, NUTS, VI, Stochastic Gradient Langevin Dynamics (SGLD)	TensorFlow	Highly customizable, integrates with Keras. Challenging for beginners. Allows for amortization.

Table 2: Comparison of Probabilistic Programming Languages

5 The Future: Directions for Marketing Research

In this section, we discuss promising areas of research and applications of PML that have, for the most part, yet to see adoption in marketing. We envision rich opportunities for marketing applications in these domains and hope to encourage the field to explore further.

5.1 Representation Learning

Extracting simpler latent representations from complex high-dimensional data is known as *representation learning* (Bengio et al. 2013), which many ML and PML methods can be viewed as doing. As seen in Section 3.3, recent marketing research has made extensive use of latent representations. Combining those models with recent innovations from computer science and statistics has great potential to improve such representations in terms of interpretability, actionability, and usefulness for downstream marketing tasks.

The methods discussed in Section 3.3 share one of two properties: they either yield representations that capture information about observed variables in a linear manner, or learn highly nonlinear mappings from representations to outcomes that may be difficult to interpret or lack desirable mathematical structure (e.g., smoothness). Recent work has focused on structuring models and estimation procedures to be both interpretable, as in the former; and flexible, as in the latter (e.g., Oblander 2023). A particularly promising direction is *causal* representation learning (Schölkopf et al. 2021), a burgeoning suite of methods that aim to extract causal (and interpretable) structures from high-dimensional data. Here, the main idea is to find a set of latent features that can be independently manipulated to change the observed data. For instance, in building a model that can generate images of cars, we would like to reduce the raw pixel space into latent features that capture product design elements (e.g., shape of the car body, headlights, wheel size) as well as features like background and angle—such settings that focus solely on reconstruction are called "unsupervised" learning. These models could help marketers understand dimensions of existing product designs and manipulate those dimensions to ideate about new ones (Sisodia et al. 2023). If we were also interested in predicting some outcome like aesthetic ratings (Burnap et al. 2023) in a "supervised" context, the causal features of interest should then ignore those that do not impact the outcome, like the photo background. These models could help marketers address design gaps

in the market (Burnap and Hauser 2018).

In practice, how does one learn causal representations? In the supervised setting, Arjovsky et al. (2019) propose combining data (e.g., photos of cars) across multiple "environments" (e.g., different dealerships), learning representations whose relationship with the outcome is invariant across environments. Wang and Jordan (2021) propose to directly optimize for causal representations using empirical bounds on counterfactual quantities. In the unsupervised setting, identifying causal representations is especially difficult. However, desirable properties of the learned representations can be enforced such as statistical independence across dimensions of the representation space (Chen et al. 2018; Khemakhem et al. 2020), independently bounded support across dimensions (Wang and Jordan 2021; Ahuja et al. 2023), or sparsity in the mapping from representations to dimensions of the observed variable (Moran et al. 2022). Under certain assumptions, the representations inferred by these approaches can recover the true causal structure of the DGP. In a related vein, Aridor et al. (2024) propose combining VAEs with Bayesian decision theory, augmenting the VAE objective function to incentivize learning representations that capture information most useful for a downstream decision task.

5.2 Causal Inference

Causal inference is crucial for understanding the effectiveness of marketing actions (e.g., paid search ads), without being misled by confounding factors (e.g., purchase intent). PML can facilitate causal inference on more complex models and datasets by leveraging the ideas discussed in Section 3, but to date, has seen limited application in marketing. In this subsection, we overview two perspectives on causal inference and discuss their potential for marketing. The Bayesian perspective on potential outcomes provides a natural way of modeling missingness in potential (or counterfactual) outcomes using latent variables. Causal graphical models help provide principled identification strategies, especially for models with nontrivial dependency structures.

Bayesian Perspective on Potential Outcomes In the potential outcomes framework (Imbens and Rubin 2015), for each unit i, there are typically four quantities of interest: the potential outcomes $Y_i(0)$ and $Y_i(1)$, treatment W_i, and covariates $\mathbf{X}_i$. Bayesian causal inference (see Li et al. 2023 for a review) treats these quantities as random variables, builds a model for them, and then

derives posterior inference. The Bayesian perspective has many potential benefits. First, even for complex models, it permits straightforward inference of any causal estimand (e.g., conditional or individual treatment effects) via marginalization or imputation of missing values, yielding valid uncertainty quantification, even in finite samples. Second, flexibly estimating treatment effects and their heterogeneity benefits from Bayesian nonparametric methods. For example, Dirichlet process models have been used for causal mediation analysis, dynamic treatment regimes, and selection correction (Oganisian and Roy 2021). Gaussian process models excel when the treatment and control response surfaces are complex, and the empirical setting exhibits high selectivity (Alaa and Van Der Schaar 2017). Bayesian additive regression trees (BART) are especially popular, given their computational scalability, ease of hyperparameter tuning, and off-the-shelf heterogeneous treatment effect estimation (Hahn et al. 2020). Third, fusing PML with causal inference provides principled ways to respect data structures common in marketing (e.g., panel data, dynamics, disparate data sources) and high-dimensionality (e.g., unstructured data like product descriptions and images, from which covariates could be discovered jointly with causal inference), as discussed extensively in Section 3. Fourth, this perspective enables integrating causal inference with Bayesian decision theory—e.g., for dynamic decision-making contexts like personalized treatment regimes for targeted promotions or recommendations, while accounting for costs.

Despite their potential benefits, these methods have seen almost no adoption in marketing. A few exceptions include Kim et al. (2020), who propose a Bayesian synthetic control framework that outperforms frequentist counterparts, especially in settings with "large p, small n" and many irrelevant covariates.[20] In the context of news headlines, Huang and Tian (2024) use a hierarchical model to infer unobserved heterogeneity in treatment effects from A/B tests with only aggregated impression and click results.

Causal Graphical Models Directed acyclic graphs (DAGs), as introduced in Section 2.1, play a central role in "structural causal models" (Pearl et al. 2016). Namely, DAGs in which edges denote direct causal relationships (which are probabilistic and thus mesh well with Bayesian machinery) are called *causal graphical models* (CGMs). CGMs can be potent tools for revealing identification

[20] Recently, Pang et al. (2022) unify synthetic control methods for temporal and staggered treatment adoption through a Bayesian posterior predictive lens based on matrix completion (of the missing counterfactual trend) under causal constraints, while enabling scalable estimation of ITEs.

strategies and reasoning about interventions. Consider the very simple CGM in Figure 2, in which a consumer's purchase intent Z is confounding the effect of ad spending W on sales Y. Using operations on the graph, one can identify which variables (e.g., Z) to condition on to identify causal effects (e.g., of W on Y). While doing so is fairly trivial in the example above, CGMs can be particularly helpful for more complex models, involving intricate patterns of mediation (e.g., $Z \to W \to Y$); mutual dependence (e.g., $Z \to W$ and $Z \to Y$); and mutual causation (e.g., $Z \to Y$ and $W \to Y$).[21]

In marketing, CGMs have been used to develop novel identification strategies: e.g., for mediation analysis with measurement error (Laghaie and Otter 2023) and for assessing the impact of reputation on persuasion with textual confounders (Manzoor et al. 2023). However, the full potential of CGMs has not been realized. We envision opportunities to identify and estimate causal quantities in complex models using structural causal modeling (e.g., where features derived from unstructured data like ad copies or user-generated content are used as treatment to assess their impact on sales; see Feder et al. 2022 for a review of methods).

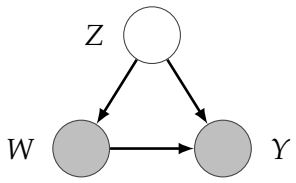

Figure 2: Simple illustrative CGM
Shaded nodes indicate observed variables. Example drawn from Blake et al. (2015).

5.3 Better Experiments and Decisions

In Section 2.4, we outlined how Bayesian methods and, by extension, PML, connect with proper uncertainty quantification and optimal decision-making. Despite this promise, there has been minimal work in marketing using PML in conjunction with decision theory. We delineate two open areas of research: (1) augmenting traditional decision-making tools with unstructured data and (2) designing better experiments.

[21] Potential outcomes also relate to this perspective. The framework provide a principled way to select the covariates that can satisfy unconfoundedness. Overlap manifests in assuming that the edges in the CGM are not deterministic. SUTVA is implicitly assumed in the graph: no interference (i.e., between units) manifests since the arrow from treatment to outcome is for each individual only; and consistency (i.e., no hidden variations of treatment) manifests in the definition of the treatment node.

Augmenting Decision-making with Unstructured Data Many decisions in marketing involve the design of creatives. One potential benefit of PML lies in the seamless incorporation of multiple information streams in making these decisions: e.g., melding text, audio, video—as well as their interactions—and determining their "optimal" content, sequencing over time, and degree of heterogeneity across target consumer groups. Doing so requires integrating unstructured data with models of consumers. Most existing models using unstructured data in marketing take a two-step approach to this task: first extracting some features from unstructured data, and then using them in a model. Treating these two steps as independent ignores the uncertainty of the initial extraction stage (e.g., Allon et al. 2023). The probabilistic perspective offers an alternative: train a model of choice jointly with a model of unstructured data, so that the features extracted are exactly those that are relevant for choice or some downstream decision (as in Section 5.1). To implement such solutions, we need better methods like Bayesian neural networks (see Section 3.2); transfer learning; and meta-learning, which often involves using hierarchical models and empirical Bayesian methods to learn priors and effectively initialize neural networks (Grant et al. 2018; Yin et al. 2019).[22]

Optimal and Adaptive Experimentation The goal of optimal experimental design is to design an experiment that maximally reduces uncertainty, in expectation, about a decision-relevant quantity, given constraints on the size of the experiment. Relatedly, in optimal *sequential* design, the goal is to optimize the design of each trial to maximize the expected usefulness of its potential results. By capturing the full posterior predictive distribution over an outcome of interest, PML enables quantifying both the optimal action for achieving a high expected value of the outcome *and also* the experimental design that will yield the maximum expected information about the problem.

We view three contexts as particularly amenable for PML approaches. First, for preference measurement, conjoint-based experimental design techniques have a long history in marketing (Green and Rao 1971; Netzer et al. 2008), for which both sequential design and Bayesian methods have been tremendously popular. PML offers new possibilities for designing such studies, especially with regards to incorporating unstructured data (Sisodia et al. 2023) as well as relaxing as-

[22] Indeed, Yin et al. (2024) leverage meta-learning to offer solutions to the cold start problem and infer time-varying, heterogeneous preferences using transformers.

sumptions about utility functions, which are likely heterogeneous and nonlinear (Dew 2023). Second, for A/B/n tests, some extant work (e.g., Schwartz et al. 2017; Feit and Berman 2019; Aramayo et al. 2023) utilize Bayesian methods for optimal decision-making. PML, in its ability to synthesize massive and potentially unstructured data, while providing posterior uncertainty quantification, can offer a solution for otherwise prohibitively large design spaces. Indeed, Campbell and Daviet (2023) investigate optimally designing a complex stimulus, like a banner ad, to maximize an objective, like click-through rate, using a combination of sequential testing and Bayesian deep learning. Third, PML methods can help design better experiments by generating the creatives themselves. For instance, Luo and Toubia (2024) leverage disentangled representations from GANs to manipulate realistic visual stimuli on a single attribute at a time. These stimuli can be used in experiments to assess the impact of those latent features on preferences (e.g., via conjoint; Sisodia et al. 2023) or other outcomes of interest.

5.4 Integration with Theory-Based Behavioral Models

Marketing researchers are often interested in estimating theory-based models of human or firm behavior, such as structural models based on microeconomic theory and other cognitive models based on psychometric or neuroscientific theory. Historically, structural models have been differentiated from ML approaches. While ML models excel at predicting decision makers' behavior conditional on a joint distribution of *observed* variables, they fail to generalize to counterfactual situations that are beyond the support of the observed data (Iskhakov et al. 2020). Structural models, on the other hand, learn "invariant" parameters governing behavior that allow for extrapolation outside the support of the training data. The inferred parameters and counterfactual simulations can then be used to interpret preferences and derive policy-relevant implications. Some such models integrate the Bayesian perspective (e.g., Bayesian updating in learning models; Ching et al. 2013), and increasingly recently, the *PML* perspective. We will highlight three such promising integrations as opportunities for further research.

Flexibly Specifying Behavioral Models Structural models often make restrictive functional form assumptions (e.g., linear utilities) which limit their explanatory power and external validity. These assumptions, which reduce computational complexity, come at the expense of inaccurately

representing complex relationships. As outlined in Section 3.2, PML approaches can be useful for capturing unknown, non-linear utility functions, but remain largely underutilized in structural modeling (e.g., Korganbekova and Zuber 2023). Another form of flexibility is properly accounting for heterogeneity across customers and products, and over time, which can result in more accurate inferences about economically relevant quantities, e.g., via Bayesian nonparametrics (Onzo and Ansari 2024) and matrix factorization (Donnelly et al. 2024). As seen by these recent advancements, PML holds significant potential to enhance how structural models account for multiple sources of heterogeneity.

Estimating Intractable Behavioral Models Estimating structural models often involves solving a complex sequence of computations, such as integrating over high dimensional state spaces to calculate the value function in dynamic discrete choice models or integrating over truncated posteriors of latent variables to calculate choice probabilities in multivariate probit models (Geweke and Keane 2001).[23] ML methods have great potential to allow for more flexible and computationally efficient approximations in these contexts. For instance, Maliar et al. (2021) approximate the value function and policy function in a dynamic program with neural networks and use stochastic gradient optimization of the Bellman equation to jointly estimate both neural networks. Scheidegger and Bilionis (2019) reduce the state space's effective dimensionality in a dynamic programming problem by projecting onto a lower dimensional surface and then using Gaussian processes to approximate the value function and policy function in this subspace. In these contexts, modern PML inference techniques (e.g., variational inference, reparameterization gradients, amortization) may be particularly useful, especially for heterogeneous models requiring solutions for many different individual sets of parameters simultaneously.

PML as Approximate Inference A conceptual limitation of many structural models is that decision-makers are assumed to behave rationally, solving optimization problems that are difficult even for researchers with powerful computers. In such contexts, it seems cognitively implausible that consumers would be able to perfectly calculate and optimize such complex objective functions. Much work in cognitive modeling treats decision-makers as processing information

[23] This issue is especially problematic when the computations depend on the unknown parameters being estimated, since every new evaluation of the likelihood involves re-solving the computation with the new parameter vector.

imperfectly: e.g., in "rational inattention" models, agents process information noisily but can select how accurate to make their computations subject to a cognitive cost of precision (Turlo et al. 2023). Consumers may make decisions using similar approximations and simplifications as those used by approximate computational techniques in PML. For example, Lin et al. (2015) propose that consumers could be learning using a bandit algorithm. Aridor et al. (2024) propose that players of an economic game adaptively learn noisy encodings of environmental information according to a VAE. More broadly, a large literature in theoretical neuroscience on the "free energy principle" posits that humans learn about unknown state variables using variational inference and select actions that maximize the ELBO of future observations (Friston 2010; Gershman 2019). Applying these ideas to model consumer behavior can not only simplify implementation, but also allow for more realistic models of how consumer decisions deviate from rationality, leading to more accurate inferences and policy implications.

6 Conclusion

In this paper, we have described popular models and methods from PML that have expanded and enriched our ability to model consumers and their choices. We highlighted how PML achieves flexibility, scalability, interpretability, and proper uncertainty quantification, all of which are crucial for making good marketing decisions. Though PML has already been influential in the field of marketing, recent advances present new and exciting opportunities, allowing for more complex data, more sophisticated models, and deeper understandings of consumer behavior. We hope this paper can both serve as an entry point for researchers interested in this space, and provide the scaffolding for future PML researchers in marketing to build upon.

References

Ahuja, K., Mahajan, D., Wang, Y., and Bengio, Y. (2023). Interventional causal representation learning. In *International Conference on Machine Learning*, pages 372–407.

Alaa, A. M. and Van Der Schaar, M. (2017). Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, volume 30.

Allenby, G. M., Arora, N., and Ginter, J. L. (1998). On the heterogeneity of demand. *Journal of Marketing Research*, 35(3):384–389.

Allon, G., Chen, D., Jiang, Z., and Zhang, D. (2023). Machine learning and prediction errors in causal inference. *Available at SSRN 4480696*.

Anand, P. and Lee, C. (2023). Using deep learning to overcome privacy and scalability issues in customer data transfer. *Marketing Science*, 42(1):189–207.

Ansari, A. and Iyengar, R. (2006). Semiparametric thurstonian models for recurrent choices: A bayesian analysis. *Psychometrika*, 71:631–657.

Ansari, A., Li, Y., and Zhang, J. Z. (2018). Probabilistic topic model for hybrid recommender systems: A stochastic variational bayesian approach. *Marketing Science*, 37(6):987–1008.

Ansari, A. and Mela, C. F. (2003). E-customization e-customization. *Journal of Marketing Research*, 40(2):131–145.

Aramayo, N., Schiappacasse, M., and Goic, M. (2023). A multiarmed bandit approach for house ads recommendations. *Marketing Science*, 42(2):271–292.

Aridor, G., da Silveira, R. A., and Woodford, M. (2024). Information-constrained coordination of economic behavior. Technical report, National Bureau of Economic Research.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv Preprint arXiv:1907.02893*.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Betancourt, M. and Girolami, M. (2015). Hamiltonian monte carlo for hierarchical models. *Current Trends in Bayesian Methodology with Applications*, 79(30):2–4.

Blake, T., Nosko, C., and Tadelis, S. (2015). Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica Econometrica*, 83(1):155–174.

Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Boughanmi, K. and Ansari, A. (2021). Dynamics of musical success: A machine learning approach for multimedia data fusion. *Journal of Marketing Research*, 58(6):1034–1057.

Boughanmi, K., Ansari, A., and Li, Y. (2023). A generative model of consumer collections. *Available at SSRN 4261182*.

Box, G. E. and Hunter, W. G. (1962). A useful method for model-building. *Technometrics*, 4(3):301–318.

Braun, M. and Bonfrer, A. (2011). Scalable inference of customer similarities from interactions data using dirichlet processes. *Marketing Science*, 30(3):513–531.

Braun, M., Fader, P. S., Bradlow, E. T., and Kunreuther, H. (2006). Modeling the "pseudodeductible" in homeowners' insurance. *Management Science*, 52(8):1258–1272.

Braun, M. and McAuliffe, J. (2010). Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335.

Burnap, A. and Hauser, J. (2018). Predicting "design gaps" in the market: Deep consumer choice models under probabilistic design constraints. *arXiv Preprint arXiv:1812.11067*.

Burnap, A., Hauser, J. R., and Timoshenko, A. (2023). Product aesthetic design: A machine learning augmentation. *Marketing Science*.

Büschken, J. and Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *marketing Science*, 35(6):953–975.

Campbell, C. and Daviet, R. (2023). Creating effective digital ads: Automatic bayesian combinatorial design. *Working Paper*.

Canny, J. (2004). Gap: a factor model for discrete data. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 122–129. ACM.

Chen, F., Liu, X., Proserpio, D., and Troncoso, I. (2022). Product2vec: Leveraging representation learning to model consumer product choice in large assortments. *Nyu Stern School of Business*.

Chen, R. T., Li, X., Grosse, R. B., and Duvenaud, D. K. (2018). Isolating sources of disentanglement in variational autoencoders. *Advances in Neural Information Processing Systems*, 31.

Cheng, Z., Lee, D., and Tambe, P. (2022). InnoVAE: Generative ai for understanding patents and innovation. *Available at SSRN 3868599*.

Ching, A. T., Erdem, T., and Keane, M. P. (2013). Learning models: An assessment of progress, challenges, and new developments. *Marketing Science*, 32(6):913–938.

Daviet, R. (2020). Bayesian deep learning for small datasets. *Working Paper*.

De Bruyn, A. and Otter, T. (2022). Bayesian consumer profiling: How to estimate consumer characteristics from aggregate data. *Journal of Marketing Research*, 59(4):755–774.

Dew, R. (2023). Adaptive preference measurement with unstructured data. *Available at SSRN 4641773*.

Dew, R. and Ansari, A. (2018). Bayesian nonparametric customer base analysis with model-based visualizations. *Marketing Science*, 37(2):216–235.

Dew, R., Ansari, A., and Li, Y. (2020). Modeling dynamic heterogeneity using gaussian processes. *Journal of Marketing Research*, 57(1):55–77.

Dew, R., Ansari, A., and Toubia, O. (2022). Letting logos speak: Leveraging multiview representation learning for data-driven branding and logo design. *Marketing Science*, 41(2):401–425.

Dew, R., Ascarza, E., Netzer, O., and Sicherman, N. (2024). Detecting routines: Applications to ridesharing customer relationship management. *Journal of Marketing Research*, 61(2):368–392.

Donnelly, R., Kanodia, A., and Morozov, I. (2024). Welfare effects of personalized rankings. *Marketing Science*, 43(1):92–113.

Donnelly, R., Ruiz, F. J., Blei, D., and Athey, S. (2021). Counterfactual inference for consumer choice across many product categories. *Quantitative Marketing and Economics*, pages 1–39.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid monte carlo. *Physics Letters B*, 195(2):216–222.

Dubé, J.-P., Hitsch, G. J., and Rossi, P. E. (2010). State dependence and alternative explanations for consumer inertia. *The Rand Journal of Economics*, 41(3):417–445.

Dzyabura, D. and Hauser, J. R. (2011). Active machine learning for consideration heuristics. *Marketing Science*, 30(5):801–819.

Farrell, M. H., Liang, T., and Misra, S. (2020). Deep learning for individual heterogeneity: An automatic inference framework. *arXiv Preprint arXiv:2010.14694*.

Feder, A., Keith, K. A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M. E., et al. (2022). Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Feit, E. M., Beltramo, M. A., and Feinberg, F. M. (2010). Reality check: Combining choice experiments with market data to estimate the importance of product attributes. *Management Science*, 56(5):785–800.

Feit, E. M. and Berman, R. (2019). Test & roll: Profit-maximizing a/b tests. *Marketing Science*, 38(6):1038–1058.

Feit, E. M., Wang, P., Bradlow, E. T., and Fader, P. S. (2013). Fusing aggregate and disaggregate data with an application to multiplatform media consumption. *Journal of Marketing Research*, 50(3):348–364.

Feng, X., Zhang, S., Srinivasan, K., et al. (2023). *Marketing Through the Machine's Eyes: Image Analytics and Interpretability*, volume 20. Emerald Publishing Limited.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Fong, H., Kumar, V., and Sudhir, K. (2023). A theory-based interpretable deep learning architecture for music emotion. *Available at SSRN 4025386*.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Gershman, S. J. (2019). What does the free energy principle tell us about the brain? *Neurons, Behavior, Data Analysis, and Theory*, 2(3):1–10.

Geweke, J. and Keane, M. (2001). Computationally intensive methods for integration in econometrics. In *Handbook of econometrics*, volume 5, pages 3463–3568. Elsevier.

Gilula, Z., McCulloch, R. E., and Rossi, P. E. (2006). A direct approach to data fusion. *Journal of Marketing Research*, 43(1):73–83.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Gopalan, P., Hofman, J. M., and Blei, D. M. (2015). Scalable recommendation with hierarchical poisson factorization. In *UAI*, pages 326–335.

Grant, E., Finn, C., Levine, S., Darrell, T., and Griffiths, T. (2018). Recasting gradient-based meta-learning as hierarchical bayes. *arXiv Preprint arXiv:1801.08930*.

Green, P. E. and Rao, V. R. (1971). Conjoint measurement-for quantifying judgmental data. *Journal of Marketing Research*, 8(3):355–363.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika Biometrika*, 57(1):97–109.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Huang, M. and Tian, L. (2024). Learning preference heterogeneity from aggregate-response online experiments. *Working Paper*.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: an Introduction*. Cambridge University Press, New York.

Iskhakov, F., Rust, J., and Schjerning, B. (2020). Machine learning and structural econometrics: Contrasts and synergies. *The Econometrics Journal*, 23(3):S81–S124.

Jacobs, B., Fok, D., and Donkers, B. (2021). Understanding large-scale dynamic purchase behavior. *Marketing Science*, 40(5):844–870.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

Kalyanam, K. and Shively, T. S. (1998). Estimating irregular pricing effects: A stochastic spline regression approach. *Journal of Marketing Research*, 35(1):16–29.

Kamakura, W. A. and Russell, G. J. (1989). A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research*, 26(4):379–390.

Karlinsky-Shichor, Y. and Netzer, O. (2023). Automating the b2b salesperson pricing decisions: A human-machine hybrid approach. *Marketing Science*.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020). Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR.

Kim, H. and Allenby, G. M. (2022). Integrating textual information into models of choice and scaled response data. *Marketing Science*, 41(4):815–830.

Kim, J. G., Menzefricke, U., and Feinberg, F. M. (2004). Assessing heterogeneity in discrete choice models using a dirichlet process prior. *Review of Marketing Science*, 2(1):1–39.

Kim, J. G., Menzefricke, U., and Feinberg, F. M. (2007). Capturing flexible heterogeneous utility curves: A bayesian spline approach. *Management Science*, 53(2):340–354.

Kim, M. and Zhang, J. (2023). Discovering online shopping preference structures in large and frequently changing store assortments. *Journal of Marketing Research*, pages 665—-686.

Kim, S., Gupta, S., and Lee, C. (2021). Managing members, donors, and member-donors for effective nonprofit fundraising. *Journal of Marketing*, 85(3):220–239.

Kim, S., Lee, C., and Gupta, S. (2020). Bayesian synthetic control methods. *Journal of Marketing Research*, 57(5):831–852.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv Preprint arXiv:1312.6114*.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer Computer*, 42(8):30–37.

Korganbekova, M. and Zuber, C. (2023). Balancing user privacy and personalization. *Working Paper*.

Kosyakova, T., Otter, T., Misra, S., and Neuerburg, C. (2020). Exact mcmc for choices from menus: Measuring substitution and complementarity among menu items. *Marketing Science*, 39(2):427–447.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*.

Laghaie, A. and Otter, T. (2023). Measuring evidence for mediation in the presence of measurement error. *Journal of Marketing Research*, 60(5):847–869.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. (2017). Deep neural networks as gaussian processes. *arXiv Preprint arXiv:1711.00165*.

Levy, S. and Montgomery, A. (2024). Relaxing functional form in choice models through gaussian processes. *Working Paper*.

Li, F., Ding, P., and Mealli, F. (2023). Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A*, 381(2247):20220153.

Li, Y. and Ansari, A. (2014). A bayesian semiparametric approach for endogeneity and heterogeneity in choice models. *Management Science*, 60(5):1161–1179.

Lin, S., Zhang, J., and Hauser, J. R. (2015). Learning from experience, simply. *Marketing Science*, 34(1):1–19.

Liu, J. and Kawaguchi, K. (2022). Segmenting consumer location-product preferences for assortment localization. *Available at SSRN*.

Liu, J. and Toubia, O. (2018). A semantic approach for estimating consumer content preferences from online search queries. *Marketing Science*, 37(6):930–952.

Liu, J., Toubia, O., and Hill, S. (2021). Content-based model of web search behavior: An application to tv show search. *Management Science*, 67(10):6378–6398.

Luo, C. (2022). Understanding diffusion models: A unified perspective. *arXiv Preprint arXiv:2208.11970*.

Luo, L. and Toubia, O. (2024). The impact of facial femininity and gender identity on perceptions and behavior: Using AI for controllable stimuli generation. Working Paper, Columbia Business School.

Maliar, L., Maliar, S., and Winant, P. (2021). Deep learning for solving dynamic economic models. *Journal of Monetary Economics*, 122:76–101.

Manzoor, E., Chen, G. H., Lee, D., and Smith, M. D. (2023). Influence via ethos: On the persuasive power of reputation in deliberation online. *Management Science*.

McCarthy, D. M. and Oblander, E. S. (2021). Scalable data fusion with selection correction: An application to customer base analysis. *Marketing Science*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.

Moran, G. E., Sridhar, D., Wang, Y., and Blei, D. (2022). Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*.

Murphy, K. P. (2022). *Probabilistic Machine Learning: An Introduction*. MIT press.

Neal, R. M. (2011). MCMC using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11).

Neal, R. M. (2012). *Bayesian Learning for Neural Networks*. Springer Science & Business Media.

Netzer, O., Toubia, O., Bradlow, E. T., Dahan, E., Evgeniou, T., Feinberg, F. M., Feit, E. M., Hui, S. K., Johnson, J., Liechty, J., and Others (2008). Beyond conjoint analysis: Advances in preference measurement. *Marketing Letters*, 19(3-4):337–354.

Oblander, S. (2023). Representation learning for behavioral analysis of complex games. *Working Paper*.

Oganisian, A. and Roy, J. A. (2021). A practical introduction to bayesian estimation of causal effects: Parametric and nonparametric approaches. *Statistics in Medicine*, 40(2):518–551.

Onzo, K. and Ansari, A. (2024). Bayesian nonparametric sequential search. *Working Paper*.

Padilla, N. and Ascarza, E. (2021). Overcoming the cold start problem of customer relationship management using a probabilistic machine learning approach. *Journal of Marketing Research*, 58(5):981–1006.

Padilla, N., Ascarza, E., and Netzer, O. (2023). The customer journey as a source of information. *Available at SSRN 4612478*.

Pang, X., Liu, L., and Xu, Y. (2022). A bayesian alternative to synthetic control for comparative case studies. *Political Analysis*, 30(2):269–288.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680.

Papamarkou, T., Skoularidou, M., Palla, K., Aitchison, L., Arbel, J., Dunson, D., Filippone, M., Fortuin, V., Hennig, P., Hubin, A., et al. (2024). Position paper: Bayesian deep learning in the age of large-scale ai. *arXiv Preprint arXiv:2402.00809*.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan kaufmann.

Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

Pitman, J. and Yor, M. (1997). The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900.

Puranam, D., Narayan, V., and Kadiyali, V. (2017). The effect of calorie posting regulation on consumer opinion: A flexible latent dirichlet allocation model with informative priors. *Marketing Science*, 36(5):726–746.

Qian, Y. and Xie, H. (2011). No customer left behind: A distribution-free bayesian approach to accounting for missing xs in marketing models. *Marketing Science*, 30(4):717–736.

Qian, Y. and Xie, H. (2014). Which brand purchasers are lost to counterfeiters? an application of new data fusion approaches. *Marketing Science*, 33(3):437–448.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822. PMLR.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286. PMLR.

Rossi, P. E. and Allenby, G. M. (2003). Bayesian statistics and marketing. *Marketing Science*, 22(3):304–328.

Rossi, P. E., McCulloch, R. E., and Allenby, G. M. (1996). The value of purchase history data in target marketing. *Marketing Science*, 15(4):321–340.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Rudolph, M., Ruiz, F., Mandt, S., and Blei, D. (2016). Exponential family embeddings. *Advances in Neural Information Processing Systems*, 29.

Ruiz, F. J. R., Athey, S., and Blei, D. M. (2020). SHOPPER: A probabilistic model of consumer choice with substitutes and complements. *The Annals of Applied Statistics*, 14(1):1–27.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature Nature*, 323(6088):533–536.

Scheidegger, S. and Bilionis, I. (2019). Machine learning for high-dimensional dynamic stochastic economies. *Journal of Computational Science*, 33:68–82.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Schwartz, E. M., Bradlow, E. T., and Fader, P. S. (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522.

Shi, W., Qu, Y., and Liu, J. (2023). Screening consumer complaints for recall management: A topic model for decision automation. *Working Paper*.

Sisodia, A., Burnap, A., and Kumar, V. (2023). Automatic discovery and generation of visual design characteristics: Application to visual conjoint. *Available at SSRN 4151019*.

Sozuer, S., Netzer, O., and Krstovski, K. (2024). A recipe for creating recipes: An ingredient embedding approach. *Available at SSRN 4686749*.

Spatharioti, S. E., Rothschild, D. M., Goldstein, D. G., and Hofman, J. M. (2023). Comparing traditional and LLM-based search for consumer choice: A randomized experiment. *arXiv Preprint arXiv:2307.03744*.

Teh, Y. W., Jordan, M., Beal, M., and Blei, D. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Tian, L. (2019). *Bayesian Nonparametrics for Marketing Response Models*. Phd thesis, University of Michigan. Available at `https://deepblue.lib.umich.edu/handle/2027.42/151696`.

Tian, L. and Feinberg, F. M. (2020). Optimizing price menus for duration discounts: A subscription selectivity field experiment. *Marketing Science*, 39(6):1181–1198.

Tian, L., Turjeman, D., and Levy, S. (2023a). Privacy preserving data fusion. *Available at SSRN 4451656*.

Tian, Z., Dew, R., and Iyengar, R. (2023b). Mega or micro? influencer selection using follower elasticity. *Journal of Marketing Research*, page 00222437231210267.

Timoshenko, A. and Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science*, 38(1):1–20.

Tirunillai, S. and Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *journal of Marketing Research*, 51(4):463–479.

Toubia, O. (2021). A poisson factorization topic model for the study of creative documents (and their summaries). *Journal of Marketing Research*, 100(71):1–17.

Toubia, O., Iyengar, G., Bunnell, R., and Lemaire, A. (2019). Extracting features of entertainment products: A guided latent dirichlet allocation approach informed by the psychology of media consumption. *Journal of Marketing Research*, 56(1):18–36.

Trusov, M., Ma, L., and Jamal, Z. (2016). Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting. *Marketing Science*, 35(3):405–426.

Turlo, S., Fina, M., Kasinger, J., Laghaie, A., and Otter, T. (2023). Discrete choice in marketing through the lens of rational inattention. *Working Paper*.

Voleti, S. and Ghosh, P. (2014). A non-parametric model of residual brand equity in hierarchical branding structures with application to us beer data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 177(1):135–152.

Wang, Y. and Blei, D. M. (2018). Frequentist consistency of variational bayes. *Journal of the American Statistical Association*.

Wang, Y. and Jordan, M. I. (2021). Desiderata for representation learning: A causal perspective. *arXiv Preprint arXiv:2109.03795*.

Wedel, M. and Zhang, J. (2004). Analyzing brand competition across subcategories. *Journal of Marketing Research*, 41(4):448–456.

Xia, F., Chatterjee, R., and May, J. H. (2019). Using conditional restricted boltzmann machines to model complex consumer shopping patterns. *Marketing Science*, 38(4):711–727.

Xiao, Y., Liang, P. P., Bhatt, U., Neiswanger, W., Salakhutdinov, R., and Morency, L.-P. (2022). Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. *arXiv Preprint arXiv:2210.04714*.

Yang, J., Zhang, J., and Zhang, Y. (2023). Engagement that sells: Influencer video advertising on tiktok. *Available at SSRN 3815124*.

Yin, M., Boughanmi, K., and Ansari, A. (2024). Meta-learning customer preference dynamics on digital platforms. *Available at SSRN 4727171*.

Yin, M., Tucker, G., Zhou, M., Levine, S., and Finn, C. (2019). Meta-learning without memorization. In *International Conference on Learning Representations*.

Zhang, L., Carpenter, B., Gelman, A., and Vehtari, A. (2022). Pathfinder: Parallel quasi-newton variational inference. *Journal of Machine Learning Research*, 23(306):1–49.

Zhong, N. and Schweidel, D. A. (2020). Capturing changes in social media content: A multiple latent changepoint topic model. *Marketing Science*, 39(4):827–846.