Constrained Bayesian Neural Network Utility in the Design of Price Promotions*

Martin Burda† Connor Campbell‡

February 24, 2025

Abstract

This paper introduces a discrete choice model with a flexible utility function in the form of a Bayesian Neural Network with consumer preference heterogeneity. Adapting Sequential Monte Carlo with Hamiltonian transitions to the model structure allows us to enforce a qualitative prior constraint on the shape of the utility function stipulating that it be non-increasing in price. We further account for model uncertainty with Bayesian Model Averaging. The predictive distribution of our model is thus obtained as a weighted average of predictive distributions of admissible neural network structures weighted by the posterior probability of each model. We apply our approach to a panel of IRI coffee purchase data that combines marketing, product attributes, and consumer demographics information. We obtain model-averaged predictive densities for own and cross price elasticity and corresponding revenue change predictions of the most popular products in a counterfactual experiment, simulating several levels of price promotion. Our framework allows managers to utilize a flexible data-driven method for understanding both the form of consumer utility and individual preference variations, as a valuable tool for making strategic pricing decisions.

Keywords: Pricing, Promotions, Consumer Utility, Consumer Choice

*We would like to thank Ron Berman, Eric T. Bradlow, Remi Daviet, Ryan Dew, Avi Goldfarb, Raghu Iyengar, Qing Liu, Yao Luo, and Yuanyuan Wan for their valuable comments and suggestions. Computations were performed on the Mist supercomputer at the SciNet HPC Consortium. SciNet is funded by Innovation, Science and Economic Development Canada; the Digital Research Alliance of Canada; the Ontario Research Fund: Research Excellence; and the University of Toronto.

†Department of Economics, University of Toronto, 150 St. George St., Toronto, ON M5S 3G7, Canada; E-mail: martin.burda@utoronto.ca

‡The Wharton School, Marketing Department, University of Pennsylvania, 3730 Walnut Street, Philadelphia, PA 19104, USA; E-mail: campcon@wharton.upenn.edu

1 Introduction

Price promotions are commonly used by managers to generate brand awareness, capture market share, or advertise their products. However, accurately predicting the financial impact of a promotion before it is run can be challenging. Consumers may respond to price changes in various ways. Understanding how they do so is necessary for effective product marketing and managerial decision-making. For example, when selecting price discounts, managers often seek to determine the number of additional units they will sell given a particular reduction in price.

A well-established tool for analyzing consumer decisions is explicitly modeling consumer utility as a function of observable product attributes, including price (Guadagni and Little, 1983) and observable demographic characteristics of consumers (see e.g. Allenby and Rossi, 2006). In practice, however, the estimation of such models can be complicated, particularly in scenarios where the analyst aims to depart from the linear functional form of utility and homogeneity of preferences across consumers.

Previous literature has found empirical evidence of non-linear preferences, due to complementarities across observable choices and the outside good (Lee et al., 2013; Kim et al., 2023) or consumer perception of the value of the outside good relative to observable products (Lee and Allenby, 2014; Kim et al., 2023). The conjoint analysis literature emphasizes that a consumer's response to changes in price depends on the levels of other attributes (Toubia, 2018; Marshall and Bradlow, 2002; McCoy et al., 2022). Consumer price sensitivity seems to interact with both observed product attributes and consumer demographics. It is also well documented in the marketing literature that consumers from different demographic categories have heterogeneous preferences over product attributes (Bradlow and Rao, 2000). Thus, preferences in many practical scenarios may be non-linear with respect to product attributes and heterogeneous across consumers.

This paper develops a novel discrete choice model to analyze the relationship between the price of a product and the probability of its purchase. In contrast to most of the existing discrete choice literature, we do not place linearity restrictions on consumer utility. Instead, we model utility via a flexible neural network, and impose a qualitative constraint that the utility function be non-increasing in price, as stipulated by economic theory (Assuncao and Meyer, 1993). Due to the Bayesian methodology underpinning our analysis, we characterize the full distribution of consumer preferences and thus allow price sensitivity to vary across consumers. The resulting non-linear qualitatively constrained model environment with consumer preference heterogeneity presents significant practical challenges for implementation. Using Sequential Monte Carlo with Hamiltonian particle transitions (Burda and Daviet, 2023), we obtain the predictive distributions of price elasticities on a test data set.

Moreover, we account for uncertainty inherent in the neural network structure via Bayesian Model Averaging (BMA), whereby the data determines the posterior weights of candidate networks with a varying number of hidden nodes that characterize different degrees of non-linearity, under the qualitative shape constraint. We apply our model to a panel of IRI coffee purchase data, combining marketing, product attributes, and consumer demographics information. We obtain model-averaged predictive densities for own and cross price elasticity and corresponding revenue change predictions in a counterfactual experiment of several levels of price promotion for the most popular products.

Many flexible methods of function approximation have been explored in statistics and the machine learning literature. However, for a function to represent a preference relation and thus be an admissible utility function, it must be continuous (Mas-Colell et al., 1995), ruling out piece-wise constant functions such as tree-based approximations. Two popular machine learning methodologies that satisfy continuity and act as nonlinear functional approximators are Bayesian Neural Networks (BNNs)[4] and nonparametric Gaussian processes (GPs).

[4]Feed-forward neural networks with one hidden layer have been shown to approximate any Borel measurable function to an arbitrary degree of accuracy (Hornik et al., 1989).

Levy (2024, ch.2) specified a GP prior on the sub-utility associated with the outside good, while modeling the utility function in a log-linear form. (Dew, 2024) formulated a GP prior on utility over low-dimensional item embeddings as opposed to the original attributes. Indeed, GPs do not scale well with input dimensions and can quickly become computationally intractable without sparse approximations (Liu et al., 2020).

Neural networks thus present an attractive flexible parametrization of consumer utility with continuous derivatives. In fact, a BNN can be thought of as a user-friendly flexible parametric representation of the nonparametric GP. In a foundational result in the study of BNNs, Neal (1996) showed that a BNN with one hidden layer converges to a GP as the number of hidden nodes grows to infinity. More recently, this result has been extended to deep architectures by showing that a BNN converges to a GP as the widths of many hidden layers are sent to infinity even if some remain finite (Agrawal et al., 2020).

Farrell et al. (2020) adopted deep neural networks for modeling parameter heterogeneity in a broad class of economic models. They focus on frequentist asymptotic inference properties of the estimated parameter functions, while our aim is Bayesian finite-sample analysis with flexible utility functions. Gabel and Timoshenko (2022) developed a neural network model for predicting customer-specific purchase probabilities in response to marketing actions. Their neural network provides a flexible functional form to approximate customer purchase behavior by linking purchase histories, frequencies, and discount coupons with purchasing probabilities. Nonetheless, the product choice within a category follows a multinomial logit model with linear utility. Moreover, unlike in our case, their model only captures purchase decisions based on loyalty card transaction data, excluding product attributes and consumer characteristics.

Current BNN implementations, such as in PyTorch and Tensorflow/Keras, typically employ Variational Inference (VI) benefiting from direct use of optimized numerical libraries developed for deterministic minimization. However, variational posteriors are only approxi-

mations to actual posteriors; the former may not accurately reflect the latter even with the number of its draws approaching infinity. Approximate optimization-based schemes such as VI can thus deliver inaccurate predictive quantities (Goan and Fookes, 2020). Benefiting from the particle scalability of a Sequential Monte Carlo sampler (Dai et al., 2022), we implement our model in real time in a substantive application with close to 3 million combined product-consumer data vectors for exact Bayesian inference avoiding VI approximations.

In the marketing literature, research on price promotions (Drechsler et al., 2017; Dawes, 2018) has analyzed the effect of different framings[5] of price promotions on consumer-level purchase probabilities. In general, a price reduction increases the probability that a consumer purchases a product, with variations in effectiveness based on framing specifics or past purchase history. The strength of this response is found to be heterogeneous across consumers. Price promotions can also influence customer retention in heterogeneous ways, with certain types of promotions being more effective in retaining new customers (Kim, 2019; Shaddy and Lee, 2020). Dubé and Misra (2023) study the welfare implications of personalized pricing with a binary logit probability model on the demand side. Nonetheless, these papers impose a linear form of consumer utility, the latter in their empirical application.

When consumers make purchasing decisions, it may often be the case that they only consider a small number of products out of the total number available for sale. This phenomenon, referred to as consideration set formation, has been studied by both marketers and economists (Roberts and Lattin, 1991; Eliaz and Spiegler, 2011). However, branded coffee, the product in our application, can be arguably thought of as a low-involvement product for consumers (Radder and Huang, 2008). Consumers may not exercise a high level of cognitive effort when making purchasing decisions related to it and thus the need for a more sophisticated consideration set formation framework is not present in our setting. We therefore opted out of consideration set formation in our framework.

[5] A "framing" refers to a particular way of describing a price promotion to consumers. A price promotion may be referred to a *discount*, *price-reduction*, *deal*, etc.

The consumer choice literature has recently expanded into the multi-product category environment. Ruiz et al. (2020) develop a discrete choice model incorporating substitutability and complementarity of various goods for consumers across different product categories. Donnelly et al. (2021) consider a scenario wherein consumers make purchasing decisions in several distinct product categories in parallel. Extending our approach to the multi-product environment presents a promising avenue for future research.

The remainder of the paper is organized as follows. In Section 2 we develop a Bayesian model-averaged neural network utility model of discrete choice with qualitative constraints. In Section 3 we apply our model to a panel data set of coffee purchases and obtain predictive densities of price elasticities. We further conduct counterfactual simulations of a price reduction and present the corresponding revenue change predictions. Section 4 concludes.

2 Model

2.1 Bayesian Neural Network Utility

Formally, our model of consumer choice is set up as follows. Consumer i chooses among $j = 1, \ldots, J$ mutually exclusive choice alternatives and one outside good ($j = 0$) at each choice occasion $t = 1, \ldots, T$. The set of alternatives is allowed to vary over time, though we omit the dependence of J on t for ease of notation. The observable attributes for each alternative j observed at the choice occasion t are collected in the $K \times 1$ vector $\boldsymbol{x}_{ijt} = (x_{ijt1}, \ldots, x_{ijtK})'$, where $\boldsymbol{x}_{ijt}$ can include attributes of the alternatives such as price or brand, characteristics of consumer i such as age or income, and their interactions. Let $\boldsymbol{x}_i = (\boldsymbol{x}'_{i11}, \ldots, \boldsymbol{x}'_{iJT})$.

At each choice occasion t, the consumer chooses the alternative y_{it} that maximizes their utility u_{ijt}. For panel data indexed by t, the vector of the chosen alternatives is denoted

$\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})'$. Consumer i's utility of alternative j is given by

$$u_{ijt} = V(\boldsymbol{x}_{ijt}) + \varepsilon_{ijt}$$

where ε_{ijt} is an idiosyncratic residual distributed extreme value type I, with the utility of the outside option normalized to zero.

We model $V(\boldsymbol{x}_{ijt})$ with a constrained BNN with P hidden neurons as follows:

$$V(\boldsymbol{x}_{ijt}, \boldsymbol{\theta}_i, P) = \theta_{0,i}^{(2)} + \sum_{p=1}^{P} \theta_{p,i}^{(2)} s(z_p(\boldsymbol{x}_{ijt}, \boldsymbol{\theta}_i^{(1)})) \tag{1}$$

$$s(z_p(\boldsymbol{x}_{ijt}, \boldsymbol{\theta}_i^{(1)})) = tan^{-1}(z_p(\boldsymbol{x}_{ijt}, \boldsymbol{\theta}_i^{(1)})) \tag{2}$$

$$z_p(\boldsymbol{x}_{ijt}, \boldsymbol{\theta}_i^{(1)}) = \theta_{p0,i}^{(1)} + \sum_{k=1}^{K} \theta_{pk,i}^{(1)} x_{ijtk} \tag{3}$$

$$\boldsymbol{\theta}_i \sim q(\boldsymbol{\theta}) \tag{4}$$

$$q(\boldsymbol{\theta}) \propto f(\boldsymbol{\theta}) \mathbf{1}(\boldsymbol{\theta} \in \mathcal{S}) + 0 \times \mathbf{1}(\boldsymbol{\theta} \in \mathcal{V}) \tag{5}$$

$$f(\boldsymbol{\theta}) \equiv MVN(\mathbf{0}, \sigma_0^2 \boldsymbol{I}) \tag{6}$$

$$P \sim \text{discrete uniform } \{1, \ldots, \mathcal{P}\} \subset \mathbb{N} \tag{7}$$

where $\mathcal{S}$ is the subset of the parameter space $\Theta \ni \boldsymbol{\theta}_i$ over which the constraint is satisfied, and $\mathcal{V}$ is the subset of Θ over which the constraint is violated, with $\mathcal{S} \cup \mathcal{V} = \Theta$ and $\mathcal{S} \cap \mathcal{V} = \emptyset$. Model complexity penalty is imposed through the specification of $f(\boldsymbol{\theta})$ in (5) that assigns a smaller joint probability to higher dimensional parameter vectors $\boldsymbol{\theta}_i$ that arise in more complex neural network models. We describe individual model components in further detail below.

The model for $V(\boldsymbol{x}_{ijt}, \boldsymbol{\theta}_i, P)$ in (1) is a special case of a feed-forward neural network structure[6] (Jospin et al., 2022), built using a succession of an input layer composed by the predictors

[6]Feed-forward networks are also referred to as Multilayer Perceptrons (MLPs) in the computer science literature.

$\boldsymbol{x}_{ijt}$, M hidden layers, and an output layer. After the initial input layer $\boldsymbol{l}_0$, each subsequent hidden layer $\boldsymbol{l}_m$ is represented as a linear transformation followed by a nonlinear "activation" function $s_m(\cdot)$. We motivate the choice of our activation function (2) in section 5.3 the Appendix. The output layer is then formed by the output of the last hidden layer:

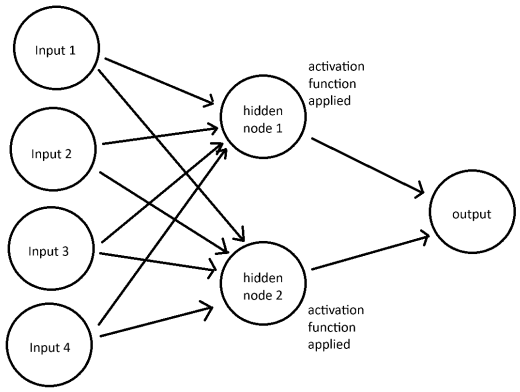
$$\text{(input layer)} \qquad \boldsymbol{l}_0 = \boldsymbol{x}_{ijt} \tag{8}$$

$$\text{(hidden layer)} \qquad \boldsymbol{l}_m = s_m\left(\theta_{m,0,i} + \boldsymbol{\theta}'_{m,i}\boldsymbol{l}_{m-1}\right), \quad \forall m \in [1, M) \tag{9}$$

$$\text{(output layer)} \qquad h(\boldsymbol{x}_{ijt}, \boldsymbol{\theta}_i) = \boldsymbol{l}_M \tag{10}$$

For each $\boldsymbol{l}_m$ the network parameters are represented by the vector $\boldsymbol{\theta}_{m,i}$ and the intercept[7] $\theta_{m,0,i}$. The complete set of the network parameters is formed by the vector $\boldsymbol{\theta}_i = (\theta_{1,0,i}, \boldsymbol{\theta}'_{1,i}, \ldots, \theta_{M,0,i}, \boldsymbol{\theta}'_{M,i})'$. A schematic diagram of a neural network model with a four-dimensional vector of inputs and two nodes in one hidden layer is presented in Figure 1. Each set of arrows represents a linear combinations of vectors of the input nodes resulting in a value of the target node.

Figure 1: A Schematic Neural Network Model



We chose to vary the model complexity by changing the number of hidden neurons P which changes the dimensionality of $\boldsymbol{\theta}_i$ while keeping one hidden layer with $M = 1$. Feed-forward neural networks with one hidden layer have been shown to approximate any Borel measurable function to an arbitrary degree of accuracy (Hornik et al., 1989).

[7]The intercept is also known as the "bias" in the ML terminology.

2.2 Shape Constraints

The model for $V(\boldsymbol{x}_{ijt}, \boldsymbol{\theta}_i, P)$ in (1) is parametrized[8] by the vector $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_i'^{(1)}, \boldsymbol{\theta}_i'^{(2)})'$ where $\boldsymbol{\theta}_i^{(1)} = \{\theta_{pki}^{(1)}, p = 1, \ldots, P, k = 0, \ldots, K\}$ and $\boldsymbol{\theta}_i^{(2)} = \{\theta_{0i}^{(2)}, \theta_{p,i}^{(2)}, p = 1, \ldots, P\}$. The constraints are imposed on the functional form of the neural network, which is enforced by a prior on the network parameters $\boldsymbol{\theta}_i$ specified in (5). The prior formulation accommodates a wide range of constraints on the BNN functional form, imposing zero posterior probability on the subset of the parameter space where the constraint is violated, $\mathcal{V}$ enforced by a "hard wall" separating $\mathcal{V}$ from its complement in the parameter space where the constraint is satisfied, $\mathcal{S}$.

In our application, we specify the constraint in terms of an inequality restriction on the first derivative of the representative utility function in (1). Thus, utility modeled by a BNN is constrained to be decreasing in price of own choice alternative, that is

$$\frac{\partial V(\boldsymbol{x}_{ijt}, \boldsymbol{\theta}_i, P)}{\partial x_{ijtk}} < 0 \text{ where } k \text{ corresponds to the price variable, for each } i, j, t. \tag{11}$$

The constraint is checked at each proposal step during our implementation algorithm. We evaluate a closed-form expression of the derivative detailed in the Appendix. If the constraint is violated at any x_{ijtk} in our data set, implying that the proposal step crosses over from $\mathcal{S}$ to $\mathcal{V}$, the proposal path is adjusted by a reflection from the constraint wall. The proposal path bounces off into $\mathcal{S}$ in a mirror image of the originally proposed move into $\mathcal{V}$, advancing the same distance as the original proposal but in the direction of the reflection. This mechanism avoids rejecting proposals because of constraint violation, leading to high numerical efficiency of the sampler.

[8] Neural network parameters are often called network "weights" in the ML literature; the term is used in the unnormalized sense as the "weights" generally do not add up to one over the network structure.

2.3 Choice Probability

Our modeling approach includes a layer of prior hierarchy over the model space that is used in the Bayesian Model Averaging (BMA) procedure for integrating out the random number of hidden neurons P, with prior given in (7). Denote the BNN utility model (1)-(6) by the short-hand notation $\mathcal{M}_P \in \mathcal{M}$, characterizing the neural network in terms of the number of hidden nodes $P = 1, \ldots, \mathcal{P}$, where $\mathcal{M}$ is the space of admissible models. Let Θ_P denote the parameter space associated with $\mathcal{M}_P$. For notational convenience we do not include the subscript P on $\boldsymbol{\theta}_i$, though the dimension of $\boldsymbol{\theta}_i$ changes with P.

Conditional on $\mathcal{M}_P$ and $\boldsymbol{\theta}_i$, the probability of consumer i choosing at t the alternative y_{it} is given by

$$q(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_i, \mathcal{M}_P) = \prod_{t=1}^{T} \frac{\exp\left(V(\boldsymbol{x}_{iy_{it}t}, \boldsymbol{\theta}_i, P)\right)}{1 + \sum_{j=1}^{J} \exp\left(V(\boldsymbol{x}_{ijt}, \boldsymbol{\theta}_i, P)\right)}. \tag{12}$$

When the uncertainty regarding the network parameters $\boldsymbol{\theta}_i$ is taken into account by marginalizing out $\boldsymbol{\theta}_i$, the probability (12) conditional on $\mathcal{M}_P$ but not conditional on $\boldsymbol{\theta}_i$ is

$$q(\boldsymbol{y}_i|\boldsymbol{x}_i, \mathcal{M}_P) = \int \prod_{t=1}^{T} \frac{\exp\left(V(\boldsymbol{x}_{iy_{it}t}, \boldsymbol{\theta}_i, P)\right)}{1 + \sum_{j=1}^{J} \exp\left(V(\boldsymbol{x}_{ijt}, \boldsymbol{\theta}_i, P)\right)} q(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i. \tag{13}$$

The functional form of the posterior is given in section 5.5 of the Appendix. The model (13) nests a traditional multinomial logit model with linear utility (McFadden and Train, 2000) as a special case for $M = 1$ and $P = 0$, with no hidden layers or nodes in the network, and activation function equal to identity. Many extensions of the multinomial logit have been proposed in the literature, with features such as nesting, latent class heterogeneity, or uncertainty regarding the choice set (for a recent literature review see e.g. Haghani et al. (2021)). Such features can also be included in our model, though the focus of the current paper is on the constrained non-linearity of the utility function. For comparability purposes, we use the special case described above as a direct benchmark of comparison in our empirical

application and refer to it as the "linear utility" model (for its mathematical description see section 5.2). In this way, we seek to assess the role that the specific feature of non-linearity of the utility function plays in the model.

2.4 Model Averaging and Prediction

The predictive distribution for a new "test" data set $\boldsymbol{y}_{N+1}, \boldsymbol{x}_{N+1}$ that has not been used in the model training is given by

$$q(\boldsymbol{y}_{N+1}|\boldsymbol{x}, \boldsymbol{y}, \mathcal{M}_P) = \int q(\boldsymbol{y}_{N+1}|\boldsymbol{x}_{N+1}, \boldsymbol{\theta}_i, \mathcal{M}_P) q(\boldsymbol{\theta}_i|\boldsymbol{y}, \boldsymbol{x}, \mathcal{M}_P) d\boldsymbol{\theta}_i. \tag{14}$$

Forming the predictive distribution requires that we marginalize $\boldsymbol{\theta}_i$ out of $q(\boldsymbol{y}_{N+1}|\boldsymbol{x}_{N+1}, \boldsymbol{\theta}_i, \mathcal{M}_P)$ by integrating with respect to the posterior distribution of $\boldsymbol{\theta}_i$. In this way, the predictive distribution (14) accounts for the uncertainty inherent in the model parameters $\boldsymbol{\theta}_i$.

However, the predictive distribution (14) is conditional on the neural network model with P hidden neurons and does not take into account the uncertainty about the model structure, such as the number of hidden neurons, which under the Bayesian paradigm is unknown and therefore random. Yet, as pointed out by Steel (2020), "it is hard to overstate the importance of model uncertainty for economic modeling". Empirical work is typically subject to a large amount of uncertainty about model specification. There are two broad strategies that have been employed in the Bayesian literature in taking model uncertainty into account: 1) Model selection that seeks to choose the "best" model out of the model space consisting of candidate models, and 2) Model averaging that uses a weighted average over the model space for inference. Model selection methods condition their inference on the chosen model and do not include the evidence contained in the alternative models, often leading to underestimation of the overall uncertainty. In contrast, model averaging accounts for a possible variation across alternative models. In the Bayesian sense, model averaging provides a natural reflection of model uncertainty.

From (14), the unconditional predictive distribution incorporating the uncertainty inherent in both the network parameters $\boldsymbol{\theta}_i$ and admissible models $\mathcal{M}_P \in \mathcal{M}$ is given by

$$q(\boldsymbol{y}_{N+1}|\boldsymbol{x}, \boldsymbol{y}) = \sum_{P=1}^{\mathcal{P}} q(\boldsymbol{y}_{N+1}|\boldsymbol{x}, \boldsymbol{y}, \mathcal{M}_P) q(\mathcal{M}_P|\boldsymbol{x}, \boldsymbol{y}) \tag{15}$$

where $q(\mathcal{M}_P|\boldsymbol{x}, \boldsymbol{y})$ is obtained as

$$q(\mathcal{M}_P|\boldsymbol{x}, \boldsymbol{y}) = \frac{q(\boldsymbol{y}|\boldsymbol{x}, \mathcal{M}_P) q(\mathcal{M}_P)}{\sum_{P=1}^{\mathcal{P}} q(\boldsymbol{y}|\boldsymbol{x}, \mathcal{M}_P) q(\mathcal{M}_P)}$$

using the marginal likelihood $q(\boldsymbol{y}|\boldsymbol{x}, \mathcal{M}_P)$ defined in (A-18) in the Appendix. The predictive distribution (15) is thus a weighted average of predictive distributions of each admissible neural network model with weights given by the relative posterior probability of each model.

In the application, we first obtain posterior draws for each $\mathcal{M}_P$ separately and then construct the non-linear utility model using the BMA equation (15). This allows us to avoid using birth and death moves associated with changing parameter space dimensionality. The subsequent numerical analysis on price elasticity and counterfactual simulation of the impact of sale prices on demand are then performed using the BMA neural network model, reflecting both parameter and model uncertainty. For details on the implementation algorithm used in the application, see section 5.4 of the Appendix.

3 Application

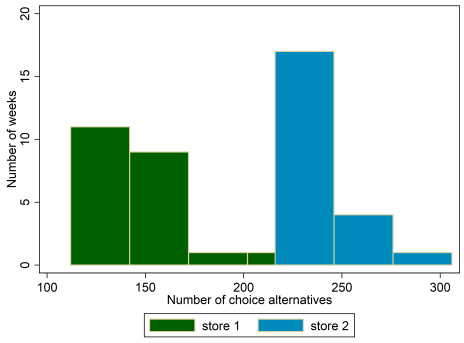
3.1 Data Description

Our empirical analysis is based on the IRI Academic Dataset (Bronnenberg et al., 2008), containing panel data of grocery store purchases in a U.S. city from June to October 2012. We chose to focus on the purchases of packaged coffee as this product category is durable,

with observed modal purchases of one package per store visit, and is generally regarded by marketers as a low-involvement good for consumers (Radder and Huang, 2008; Ahmed et al., 2004). The data set used in our analysis was created by combining IRI data files of three different types (with their respective variables): product marketing variables (price, on sale, display), product attributes (package volume, indicators for whole bean, decaffeinated, flavored, brand, store), and household demographic characteristics (income, age, education, children, married). In the empirical analysis, we use the store-consumer-product information for which the three types of data intersect. Thus, our data set contains marketing and attribute information on 408 product choice alternatives, demographic characteristics of unique 1,413 households, observed for 22 weeks for two stores on 15,413 choice occasions, totaling 2,806,226 combined product-consumer data vectors.

The IRI data set contains observations on household visits to stores in the city with purchases of a wide range of consumer products[9]. Household visits to the stores with a purchase of any of these products but not coffee then constitute the outside option, normalized to have zero utility for identification purposes. The observed set of choice alternatives varies by week and store. Figure (2) shows the variation in the choice set size in our data.

Figure 2: Packaged Coffee Choice Set Size



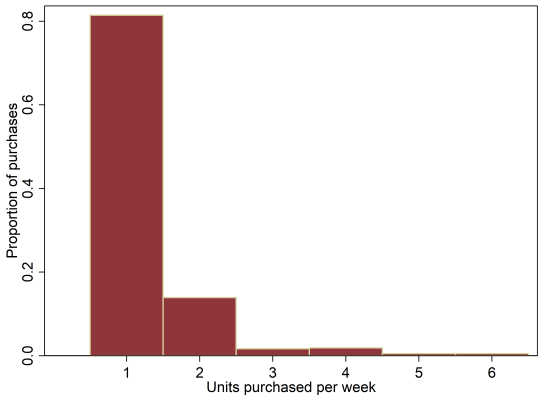
We do not observe the exact timing of the purchase. Our data set provides information only

[9]These include carbonated beverages, cold cereals, deodorants, diapers, facial tissues, frozen dinners, frozen pizza, household cleaning products, laundry detergents, margarine and butter, mayonnaise, milk, mustard and ketchup, paper towels, peanut butter, razors, salty snacks, shampoo, soup, spaghetti sauce, sugar substitutes, tissues, toothpaste, and yogurt.

about the week in which the purchase took place. However, the vast majority of consumers (81%) purchased only one unit of coffee packaging during any given week, with 14% of consumers buying two units and less than 5% of consumers getting three or more units. Hence we believe that our setup of a binary choice model with weekly choice occasions is appropriate in this context, with the data not providing sufficient variation for a count-based model or a time-of-purchase duration-based model. The relative proportions or coffee packaging units purchased by households per week are shown in Figure 3a. The variation in the amount of coffee consumed appears to be determined by the package volume as opposed to the number of packages purchased. Figure 3b presents the proportions of various package volumes (in ounces) for individual purchases in our sample.

Figure 3: Packaged Coffee Purchase Characteristics

(a) Units Purchased



(b) Unit Volume

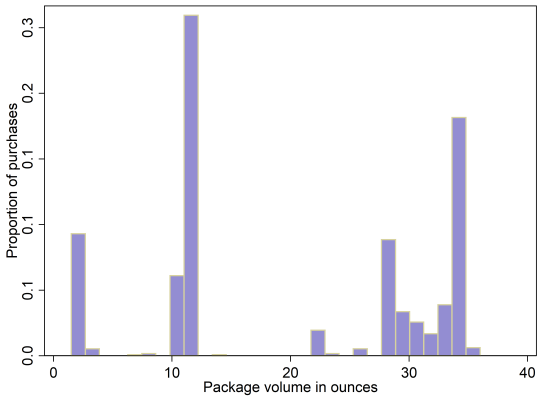


Table A1 in the Appendix provides summary statistics for the marketing variables, product attributes, and household demographic characteristics. Table A2 details the ordinal coding for the income ranges and Table A3 for the age ranges in our dataset. We include all observed product and household demographic characteristics given in tables A1, A2, and A3 in the vector of product and demographic attributes $\boldsymbol{x}_{ijt}$. Note that we also model the utility of a specific consumer i for a specific product j as functions of observable attributes specific to product j and demographics specific to consumer i.

3.2 Implementation and Model Averaging

We implemented[10] the constrained HSMC procedure described in section 5.4 of the Appendix with 100 particles for 10,000 Monte Carlo iterations retaining output from every 10th iteration, with an additional 1,000 initial iterations for burn-in. We used a relatively diffuse prior, independent Normal with mean zero and variance 100, for each model parameter. The prior embeds a penalty on model complexity with smaller implied joint distribution for higher parameter dimensions, favoring the linear utility model. The posterior parameter draws are stable and mix well.

Using the posterior draws, we tested whether a linear utility specification (see section 5.2) was supported by the data in the application against the alternative of a nonlinear utility function modeled by a neural network subject to the constraint (11) of being non-increasing in price of own choice alternative. The linear utility specification was strongly rejected with a Bayes factor of over 100, supporting the need for a flexible nonlinear utility model. Figure 4 shows the log posterior mean for the linear utility model (0 hidden neurons) and for neural network model specifications ranging from 1 to 5 hidden neurons. Figure 5 shows the trace plot of the log posterior mean for these models, displaying good mixing properties with stable means.

The neural network models with 2 and 3 hidden neurons feature overlapping posterior densities with the highest mean and in the Bayesian Model-averaged Neural Network (BMANN) (15) they carry the weights 31% and 69%, respectively, with close to zero weight assigned to the remaining neural network specifications.

Next, we compared the predictive performance of the linear utility model with the BMANN on a test data set for purchases in the four weeks of November 2012 that was not used in

[10]The implementation was run on the Unix cluster Mist at redacted (Ponce et al., 2019; Loken et al., 2010), which is comparable to many readily available commercial cloud computing services such as AWS or IBM Cloud. The full run, including output, took less than one day.

obtaining the posterior parameter draws. The log predictive densities for each model on the test set is presented in Figure 6. The BMANN model clearly dominates the linear utility model in predicting the future outcomes on the test set.

Figure 4: Log Posterior Mean

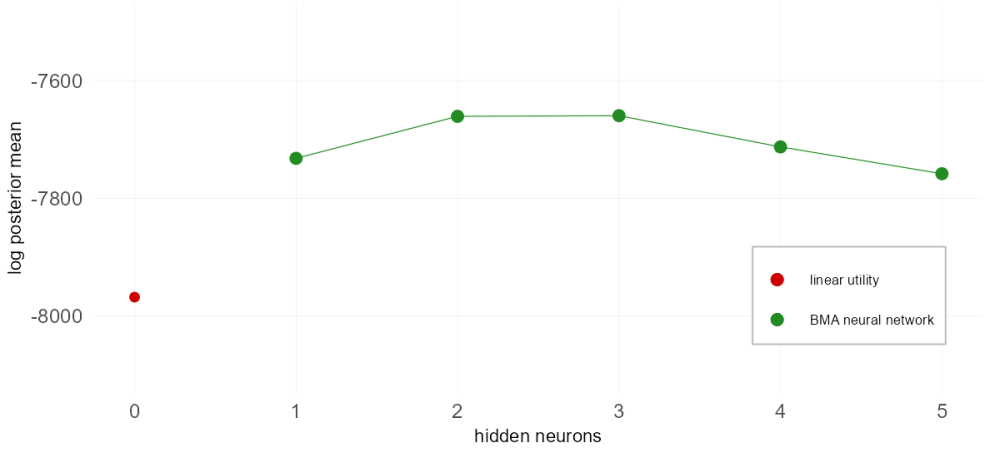


Figure 5: Log Posterior Trace Plot

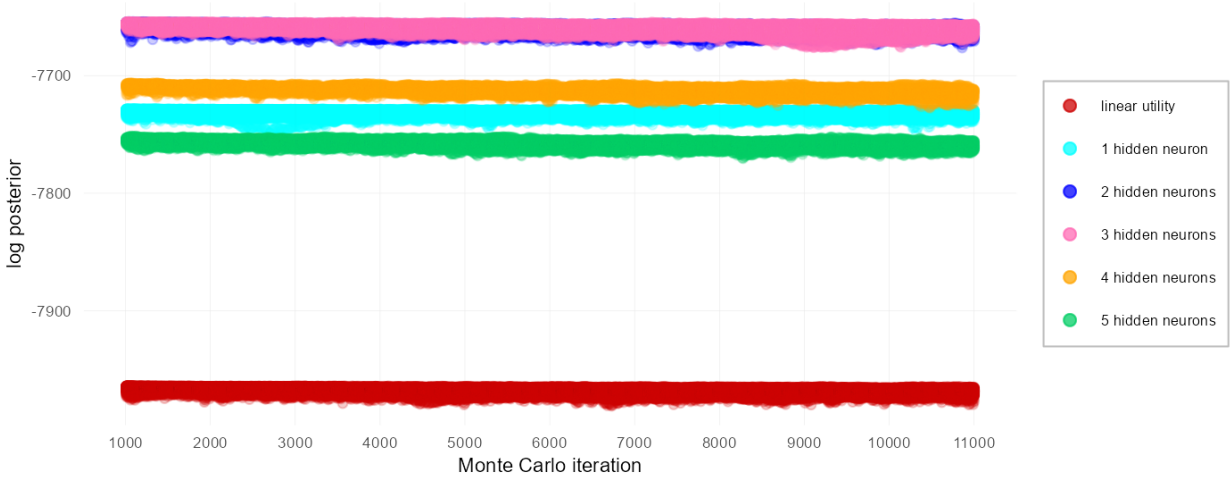
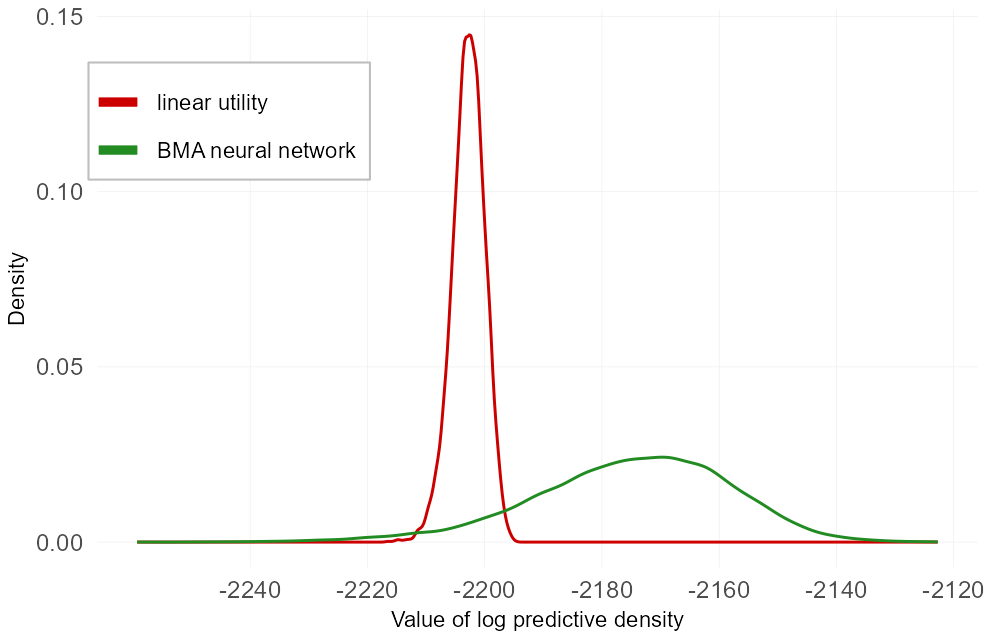


Figure 6: Log Predictive Density on a Test Set

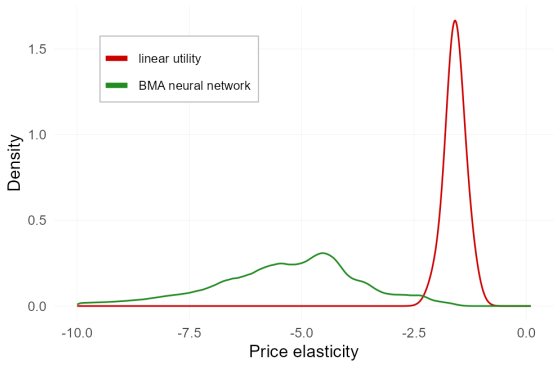


3.3 Price Elasticity

Based on the posterior draws of the parameters, we have calculated own and cross price elasticity for the most popular package in the test set in terms of number of purchases, the Folgers 34oz can of Classic Roast ground coffee, and for its key brand competitor in the same product category, the Maxwell House 31oz can of Original Roast ground coffee. The predictive densities of the price elasticities are presented in Figures 7 and 8, respectively. For both packages, the price elasticity predicted by the BMANN model is substantially higher than the one predicted by the linear utility model. This is also the case for cross-price elasticity, albeit to a smaller extent. Since the BMANN model dominates the linear utility model both in terms of model fit and predictive accuracy, we believe that a managerial decision based on the linear utility model would significantly underestimate the actual consumer purchase reaction to the product price change. We validate this notion in the next section in a counterfactual simulation of a price sale.

Figure 7: Folgers 34oz Can, Classic Roast

(a) Own Price Elasticity



(b) Cross Price Elasticity

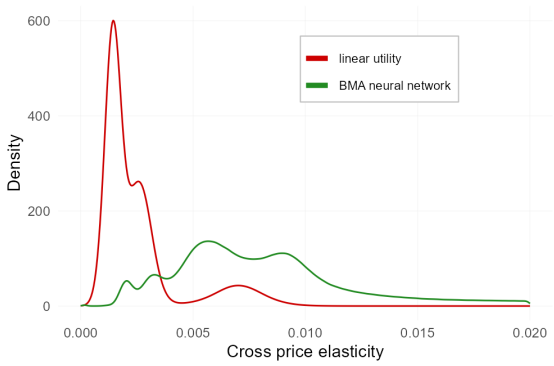
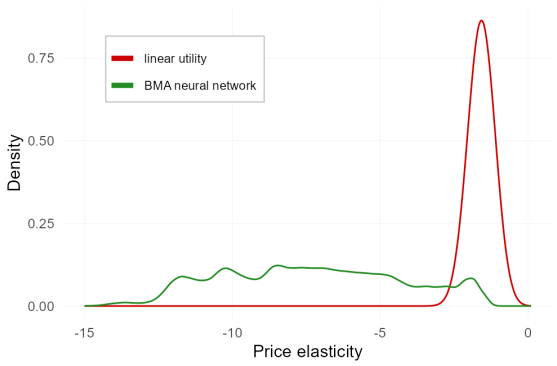
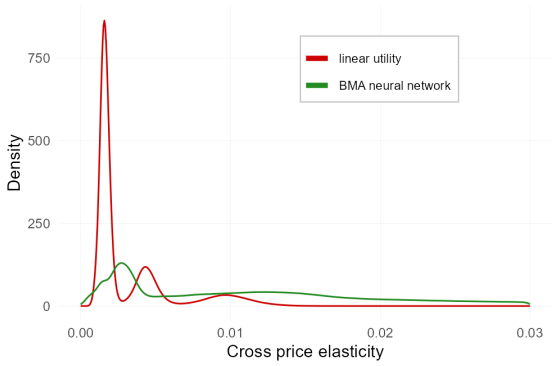


Figure 8: Maxwell House 31oz Can, Original Roast

(a) Own Price Elasticity



(b) Cross Price Elasticity



3.4 Counterfactual Simulations

Changing the price of a product in a multinomial choice scenario influences sales not only by impacting the demand for the product but also by consumer behavior substituting to or from other choice alternatives. To assess the total change in coffee revenue following an "on-sale" event we ran two counterfactual simulation scenarios by lowering the price, in turn, of each of the two packages analyzed in the elasticity section above. We first checked the mean price

change of all items in our data set corresponding to the dummy variable "on sale" equal to one as compared to its zero value: on average such change amounted to close to 15% of the original price. Consequently, in our simulations we used three simulated price changes, 10%, 15%, and 20%, around the average sale value, and also set the "on sale" dummy variable equal to one.

Table 1 shows the resulting change in revenue for the Folgers large can and Table 2 for the Maxwell House large can, using summary statistics of the predictive distribution on the test set. For both the Maxwell House large can and Folgers, the BNN model yielded substantially larger revenue increases than what was predicted by the linear utility model which is consistent with the price elasticities presented in Figures 7 and 8. The simulation was run for the test set data of one month and two grocery stores, averaging about $44 per store for a 15% reduction in price[11]. According to Statista.com, there are approximately 62 thousand stores in the USA. Naturally, these vary in size and location, but assuming that the stores in our sample are on average representative of a typical store, the BMANN model predicts a change in revenue resulting from a 15% price reduction of either package to be in the ballpark of $2.7 million[12], with the linear utility model predicting only $140 thousand[13] for Folgers and $20 thousand[14] for Maxwell House given a 15% reduction in price.

[11] For a 15% reduction, the BMANN reports a predicted change in revenue of approximately $88 per Tables 1 and 2. The approximate average change per-store is thus given by $44 = \frac{88}{2}$.

[12] Computed as $\frac{\$88.00}{2} \times 62,000 \approx \$2.7$ million.

[13] $\frac{\$4.49}{2} \times 62,000 = \approx \$140,000$.

[14] $\frac{\$0.63}{2} \times 62,000 \approx \$20,000$.

Table 1: Expected total revenue change in $USD of Folgers 34oz can, Classic Roast

Model	Price Reduction	$ Mean	$ St.Dev.	% Mean	% St.Dev.
linear utility	10%	$4.32	0.50	0.282	0.033
	15%	$4.49	0.54	0.294	0.036
	20%	$4.57	0.59	0.299	0.039
BMANN	10%	$53.27	4.95	2.759	0.224
	15%	$88.04	9.35	4.558	0.428
	20%	$141.78	16.68	7.339	0.772

Table 2: Expected total revenue change in $USD of Maxwell House 31oz can, Original Roast

Model	Price Reduction	Mean	St.Dev.	% Mean	% St.Dev.
linear utility	10%	$0.49	0.13	0.033	0.009
	15%	$0.63	0.19	0.042	0.013
	20%	$0.67	0.25	0.044	0.017
BMANN	10%	$48.42	6.43	2.511	0.287
	15%	$88.60	11.60	4.594	0.518
	20%	$143.15	18.21	7.423	0.810

4 Conclusion

In this paper, we develop a flexible consumer choice model for predicting consumer responses to price promotions. Inference via Sequential Monte Carlo with Hamiltonian particle transitions enables us to incorporate known insights from economics and marketing by imposing qualitative shape constraints on consumer preferences. We account for model uncertainty with Bayesian Model Averaging and obtain the predictive distribution of our model as a weighted average of predictive distributions of admissible neural network structures with the weights determined by the posterior probability of each model. We apply our approach to a panel of IRI coffee purchase data with information on marketing activities, product attributes, and consumer demographics. Counterfactual experiments quantify model-averaged predicted revenue change resulting from simulating several levels of price promotion. Our model enables managers to take a data-driven approach to learning the functional form of

consumer utility and heterogeneity, informing promotion pricing decisions. The implementation of our method reveals a substantially larger response from consumers to a given price promotion. We find that linear utility models underestimate the impact of a 15% price reduction by approximately $2.5 million.

Our work has some limitations that contain opportunities for future research. First, our approach requires a large amount of highly granular data on consumer-level purchases for effective neural network training and inference. Learning consumer preferences in the presence of sparse data is an area of active research within marketing (Dew, 2024). Second, interpreting the links between consumer demographic characteristics and price sensitivity within the neural network is beyond the scope of this paper but would be an interesting topic for future study. Finally, while we do not model consumers as forward-looking, which is arguably justifiable for grocery product categories including coffee, price promotions might make consumers more impatient relative to normal pricing scenarios (Shaddy and Lee, 2020). It would be interesting to empirically investigate to what extent such an effect holds for various products within our modeling framework.

Table A1: Data Summary Statistics

	Mean	SD	Min	Max
Price	9.29	3.39	3.07	28.62
Display	0.026	0.202	0	2
On sale	0.238	0.426	0	1
Income	7.475	3.103	1	12
Age	5.081	1.145	2	7
Education	3.987	1.535	0	8
Children	0.222	0.527	0	2
Married	0.096	0.295	0	1
Volume	0.867	0.581	0.094	2.250
Whole Bean	0.257	0.438	0	1
Decaf	0.088	0.284	0	1
Flavoured	0.110	0.314	0	1
Brand MaxHs	0.059	0.236	0	1
Brand Folg	0.074	0.261	0	1
Brand EightOC	0.044	0.206	0	1
Brand Strbks	0.049	0.216	0	1
Store	0.600	0.490	0	1

5 Appendix

5.1 Summary Statistics

Table A2: Income Categories

Code	*Household Income per Year*
1	below $ 9,999
2	$10,000 to $11,999
3	$12,000 to $14,999
4	$15,000 to $19,999
5	$20,000 to $24,999
6	$25,000 to $34,999
7	$35,000 to $44,999
8	$45,000 to $54,999
9	$55,000 to $64,999
10	$65,000 to $74,999
11	$75,000 to $99,999
12	above $99,999

Table A3: Age Categories

Code	*Age Range*
1	18 – 24
2	25 – 34
3	35 – 44
4	45 – 54
5	55 – 64
6	65+

5.2 Baseline Model

As a benchmark model, we define a linear utility model based largely on the logit model presented in chapter 3 of McFadden and Train (2000). Consumer i chooses among $j = 1, \ldots, J$ mutually exclusive choice alternatives and one outside good ($j = 0$) at each choice occasion $t = 1, \ldots, T$. The set of alternatives is allowed to vary over time, as is the case in the BNN model, but we omit the dependence of J on t for ease of notation. The observable attributes for each alternative j observed at the choice occasion t are defined as a $K \times 1$ vector $\boldsymbol{x}_{ijt} = (x_{ijt1}, \ldots, x_{ijtK})'$, where $\boldsymbol{x}_{ijt}$ includes all observed product and demographic attributes given in tables A1, A2, and A3. At each choice occasion t, the consumer chooses alternative y_{it} that maximizes their utility u_{ijt}. For panel data indexed by t, the observed vector of choice alternatives for consumer i over each choice occasion $t = 1, \ldots, T$ is denoted $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})'$. In our benchmark linear utility model, consumer i's utility of alternative j is given by

$$u_{ijt} = \beta_i \boldsymbol{x}_{ijt} + \epsilon_{ijt}$$

where ϵ_{ijt} is a type I extreme value distributed residual, with utility of the outside option normalized to 0. We impose a minimally informative prior distribution for the vector of parameters β_i denotes as $\boldsymbol{\beta}_i \sim p(\boldsymbol{\beta})$ where $p(\boldsymbol{\beta}) = MVN(0, \sigma_0^2 \mathbf{I})$. For our purposes, we set $\sigma_0^2 = 100$ to ensure the prior has a minimal impact on the estimation of our benchmark model. Under these given assumptions, we can obtain a closed-form expression for the choice probabilities and associated likelihood function (McFadden and Train, 2000).

Given this linear-utility specification, the conditional choice probabilities of alternative j on

choice occasion t for consumer i are given by

$$P_{ijt} = \frac{e^{\boldsymbol{\beta}_i \boldsymbol{x}_{ijt}}}{\sum_{j'=0}^{J} e^{\boldsymbol{\beta}_i \boldsymbol{x}_{ij't}}}$$

which, given the observed choice alternatives $\boldsymbol{y}_t = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_N')'$ and observed product and demographic attributes $\boldsymbol{x}_t = (\boldsymbol{x}_{11t}', \boldsymbol{x}_{12t}', \ldots, \boldsymbol{x}_{NJt}')'$ of choice alternative $j = 1, \ldots, J$ for all consumers $i = 1, \ldots, N$ yields the likelihood function

$$p(\boldsymbol{y}_t | \boldsymbol{\beta}_i, \boldsymbol{x}_t) = \prod_{t=1}^{T} \prod_{i=1}^{N} \prod_{j=1}^{J} P_{ijt}^{y_{ijt}} (1 - P_{ijt})^{1 - y_{ijt}}.$$

Thus, the posterior distribution of our benchmark model is defined as

$$p(\boldsymbol{\beta}_i | \boldsymbol{y}_t, \boldsymbol{x}_t) \propto p(y_t | \boldsymbol{\beta}_i, \boldsymbol{x}_{ijt}) p(\boldsymbol{\beta}_i). \tag{A-16}$$

5.3 Neural Network Activation Function

In (2) we use the *arctan* activation function

$$s(z) = tan^{-1}(z)$$

for two reasons. First, the function has polynomial derivatives

$$\frac{\partial s(z)}{\partial z} = (1 + z^2)^{-1}$$

that are numerically stable even for large values of z and avoids the need for parameter normalization. In contrast, the alternative typical sigmoid activation function

$$s_{\text{sig}}(z) = (1 + exp(-z))^{-1}$$

features exponential derivatives

$$\frac{\partial s_{\text{sig}}(z)}{\partial z} = \frac{1 - (1 + exp(-z))^{-1}}{(1 + exp(-z))},$$

similar to the hyperbolic tangent activation function

$$s_{tanh}(z) = \frac{exp(z) - exp(-z)}{exp(z) + exp(-z)}$$

with

$$\frac{\partial s_{tanh}(z)}{\partial z} = 1 - \left[\frac{exp(z) - exp(-z)}{exp(z) + exp(-z)}\right]^2.$$

In our experience, the exponential function in the denominator can exhibit explosive behavior and quickly lead to numerical instability for plausible argument ranges without careful normalization of the network parameters.

Second, in our application we start the Markov chains of parameter draws from a vector close to the modal value of the posterior, obtained by numerical maximization of the posterior (the maximum a-posteriori or MAP estimate). For neural network models the optimization process can take a long time without a good starting value. The general s-shape of the *arctan* function is similar to the sigmoid and *tanh* functions, but unlike in the latter two, the intercept of the *arctan* function at the y-axis passes through zero, $tan^{-1}(z = 0) = 0$.

This allows us to obtain an excellent starting value for optimization of a more complex model with P hidden nodes from the MAP estimate of a simpler model with $P-1$ hidden nodes by initializing all parameters associated with the added node P at zero, resulting in starting with $z_P = 0$ and hence $s(z_P) = tan^{-1}(z_P) = 0$, which ensures that the initial likelihood value for the optimization procedure will be at least as good as for the simpler model. In our experience, starting anywhere else typically results in a drastic reduction of the likelihood value and corresponding increase in optimization time.

5.4 Sequential Monte Carlo Implementation

There are two broad categories of implementation algorithms for BNNs: Variational Inference (VI) and Monte Carlo (MC)-based methods. VI is an approximate inference method that assumes the form of the posterior distribution, typically from a parametric family, and minimizes the Kullback-Leibler distance from the actual posterior via optimization methods used in frequentist ANNs.

The VI parametric approximation assumptions are not imposed in MC-based methods that sample from the actual posterior, though these are typically more difficult to implement. A popular numerical algorithm for BNNs in this category is Hamiltonian Monte Carlo (HMC) (Neal, 1996) that uses gradient information in posterior sampling. HMC has been shown to yield samples far more efficient than obtained by the random walk Metropolis-Hastings (RWMH) mechanism (Neal, 2011). However, HMC is inherently serial by construction, whereby a new draw of the desired parameter chain can only be taken conditional on completing the preceding draw. This imposes limitations on the numerical efficiency and scalability.

Parallel, scalable posterior sampling is enabled by Sequential Monte Carlo (SMC), also known as a particle filter (Doucet et al., 2001). SMC uses a genetic mutation-selection sampling approach with a set of particles representing the posterior distribution of a stochastic process. SMC is highly parallelizable as the core computational load involving the model likelihood is performed by individual particles independently of one another.

In this paper we use an efficient version of SMC with Hamiltonian particle transitions (Burda and Daviet, 2023). Particle values are initialized from high posterior value vectors over the subset of the parameter space where the constraint is satisfied, with $\boldsymbol{\theta} \in \mathcal{S}$. The constraint is then enforced in the HSMC mutation phase for each particle. During this phase, at iteration r of the algorithm, HSMC constructs a sequence $\{\boldsymbol{\theta}_r^\ell\}_{\ell=1}^L$ according to the Hamiltonian dynamics starting from the current state $\boldsymbol{\theta}_r^0$ and setting the last member of the sequence as the new state proposal $\boldsymbol{\theta}_{r+1}^* = \boldsymbol{\theta}_r^L$. The proposal sequence is generated using difference equations of the law of motion yielding high acceptance probability even for proposals that are relatively distant from the current draw in the parameter space. This facilitates efficient exploration of the parameter space with the resulting Markov chain (Leimkuhler and Reich, 2004).

Constraints are incorporated into the HSMC proposal mechanism via "hard walls" representing a barrier against which the proposal sequence, simulating a particle movement, bounces off elastically. Heuristically, the constraint is checked at each step of the proposal sequence and if it is violated then the trajectory of the sequence is reflected off the hard wall posed by the constraint. This facilitates efficient exploration of the parameter space even in the presence of highly complex parameter constraints (Neal, 2011). We state the functional form of the derivative of utility with respect to price in (A-21). As a sufficient condition that is fast to evaluate numerically, in our implementation the constraint is violated if $\theta_{pi}\theta_{p1i} > 0$

for any $p = 1, \ldots, P$.

Throughout the proposal path construction the Hamiltonian dynamics use posterior gradient information. In the next section we provide a closed-form expression of the gradient of our neural network model.

5.5 The Posterior and its Gradient

Let $\boldsymbol{x} = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_N')'$ and $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_N')'$. Due to the independence assumption on i, the conditional likelihood over the sample can be expressed as

$$q(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_i, \mathcal{M}_P) = \prod_{i=1}^{N} q(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_i, \mathcal{M}_P) \tag{A-17}$$

with log-likelihood

$$\ln q(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_i, \mathcal{M}_P) = \sum_{i=1}^{N}\sum_{t=1}^{T} V(\boldsymbol{x}_{iy_{it}}, \boldsymbol{\theta}_i, P) - \sum_{i=1}^{N}\sum_{t=1}^{T} \ln\left[1 + \sum_{j=1}^{J} \exp\left(V(\boldsymbol{x}_{ijt}, \boldsymbol{\theta}_i, P)\right)\right].$$

The posterior is formed as

$$q(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}, \mathcal{M}_P) = \frac{q(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_i, \mathcal{M}_P) q(\boldsymbol{\theta}_i)}{\int_{\Theta_P} q(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_i, \mathcal{M}_P) q(\boldsymbol{\theta}) d\boldsymbol{\theta}_i} = \frac{q(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_i, \mathcal{M}_P) q(\boldsymbol{\theta})}{q(\boldsymbol{y}|\boldsymbol{x}, \mathcal{M}_P)} \tag{A-18}$$

with the second equality defining $q(\boldsymbol{y}|\boldsymbol{x}, \mathcal{M}_P)$, the marginal likelihood of $\mathcal{M}_P$ evaluated in the Sequential Monte Carlo algorithm.

For a large sample of R draws $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(R)}\}$ from the posterior (A-18), (14) can be numer-

ically approximated by

$$q(\boldsymbol{y}_{N+1}|\boldsymbol{x},\boldsymbol{y},\mathcal{M}_P) \approx \frac{1}{R}\sum_{r=1}^{R} q(\boldsymbol{y}_{N+1}|\boldsymbol{x}_{N+1},\boldsymbol{\theta}_i^{(r)},\mathcal{M}_P).$$

Let θ represent a generic element of the parameter vector $\boldsymbol{\theta}$. Using (A-17) in (A-18) we obtain

$$\begin{aligned}\frac{\partial ln\left[q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}_i,\mathcal{M}_P)q(\boldsymbol{\theta}_i)\right]}{\partial\theta_i} &= \frac{\partial ln\left[\prod_{i=1}^{N} q(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_i,\mathcal{M}_P)q(\boldsymbol{\theta}_i)\right]}{\partial\theta_i} \\ &= \frac{\partial\left[\sum_{i=1}^{N} ln\left[q(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_i,\mathcal{M}_P)\right] + ln\left[q(\boldsymbol{\theta}_i)\right]\right]}{\partial\theta} \\ &= \sum_{i=1}^{N}\frac{\partial ln\left[q(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_i,\mathcal{M}_P)\right]}{\partial\theta_i} + \frac{\partial ln\left[q(\boldsymbol{\theta}_i)\right]}{\partial\theta_i} \qquad \text{(A-19)}\end{aligned}$$

Using (12) in (A-19) yields

$$\begin{aligned}\frac{\partial ln\left[q(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_i,\mathcal{M}_P)\right]}{\partial\theta} &= \sum_{t=1}^{T}\frac{\partial ln\left[\frac{\exp\left(V(\boldsymbol{x}_{iy_{it}t},\boldsymbol{\theta}_i,P)\right)}{1+\sum_{j=1}^{J}\exp(V(\boldsymbol{x}_{ijt},\boldsymbol{\theta}_i,P))}\right]}{\partial\theta} \\ &= \sum_{t=1}^{T}\frac{\partial V(\boldsymbol{x}_{iy_{it}t},\boldsymbol{\theta}_i,P)}{\partial\theta} - \sum_{t=1}^{T}\frac{\partial ln\left[1+\sum_{j=1}^{J}\exp\left(V(\boldsymbol{x}_{ijt},\boldsymbol{\theta}_i,P)\right)\right]}{\partial\theta} \\ &= \sum_{t=1}^{T}\frac{\partial V(\boldsymbol{x}_{iy_{it}t},\boldsymbol{\theta}_i,P)}{\partial\theta_i} - \sum_{t=1}^{T}\left[1+\sum_{j=1}^{J}\exp\left(V(\boldsymbol{x}_{ijt},\boldsymbol{\theta}_i,P)\right)\right]^{-1} \\ &\quad\times\exp\left(V(\boldsymbol{x}_{ijt},\boldsymbol{\theta}_i,P)\right)\frac{\partial V(\boldsymbol{x}_{ijt},\boldsymbol{\theta}_i,P)}{\partial\theta} \qquad \text{(A-20)}\end{aligned}$$

We can now evaluate (A-20) using functional form specifications for the network, $V(\cdot)$, the

prior, $q(\boldsymbol{\theta})$, and for specific elements of the parameter vector $\boldsymbol{\theta}$. Using (3) in (2),

$$V(\boldsymbol{x}_{ijt}, \boldsymbol{\theta}_i, P) = \theta_{0,i} + \sum_{p=1}^{P} \theta_{pi} s \left(\theta_{p0i} + \sum_{k=1}^{K} \theta_{pki} x_{ijtk} \right)$$

The derivative in the first and second RHS term of (A-20) then become

$$\begin{aligned}
\frac{\partial V(\boldsymbol{x}_{ijt}, \boldsymbol{\theta}_i, P)}{\partial \theta_{p0i}} &= \theta_{pi} \frac{\partial s(z_p(\boldsymbol{x}_{ijt}))}{\partial \theta_{p0i}} = \theta_{pi} \left(1 + z_p(\boldsymbol{x}_{ijt})^2\right)^{-1} \\
\frac{\partial V(\boldsymbol{x}_{ijt}, \boldsymbol{\theta}_i, P)}{\partial \theta_{pki}} &= \theta_{pi} \frac{\partial s(z_p(\boldsymbol{x}_{ijt}))}{\partial \theta_{pki}} = \theta_{p,i} \left(1 + z_p(\boldsymbol{x}_{ijt})^2\right)^{-1} x_{ijtk} \\
\frac{\partial V(\boldsymbol{x}_{ijt}, \boldsymbol{\theta}_i, P)}{\partial \theta_0} &= 1 \\
\frac{\partial V(\boldsymbol{x}_{ijt}, \boldsymbol{\theta}_i, P)}{\partial \theta_{p,i}} &= s(z_p(\boldsymbol{x}_{ijt}))
\end{aligned}$$

For enforcing the constraints in our application, the derivative takes the form

$$\begin{aligned}
\frac{\partial V(\boldsymbol{x}_{ijt}, \boldsymbol{\theta}_i, P)}{\partial x_{ijt1}} &= \sum_{p=1}^{P} \theta_{pi} \frac{\partial s(z_p(\boldsymbol{x}_{ijt}))}{\partial z_p(\boldsymbol{x}_{ijt})} \frac{\partial z_p(\boldsymbol{x}_{ijt})}{\partial x_{ijt1}} \\
&= \sum_{p=1}^{P} \theta_{pi} \left(1 + z_p(\boldsymbol{x}_{ijt})^2\right)^{-1} \theta_{p1i}
\end{aligned} \tag{A-21}$$

where

$$z_p(\boldsymbol{x}_{ijt}) = \theta_{p0i} + \sum_{k=1}^{K} \theta_{pki} x_{ijtk}.$$

References

Agrawal, D., T. Papamarkou, and J. Hinkle (2020). Wide neural networks with bottlenecks are deep gaussian processes. *Journal of Machine Learning Research 21*(175), 1–66.

Ahmed, Z. U., J. P. Johnson, X. Yang, C. Kheng Fatt, H. Sack Teng, and L. Chee Boon (2004). Does country of origin matter for low-involvement products? *International marketing review 21*(1), 102–120.

Allenby, G. M. and P. E. Rossi (2006). Hierarchical bayes models. *The handbook of marketing research: Uses, misuses, and future advances*, 418–440.

Assuncao, J. L. and R. J. Meyer (1993, May). The Rational Effect of Price Promotions on Sales and Consumption. *Management Science 39*(5), 517–535. Publisher: INFORMS.

Bradlow, E. T. and V. R. Rao (2000). A hierarchical bayes model for assortment choice. *Journal of Marketing Research 37*(2), 259–268.

Bronnenberg, B. J., M. W. Kruger, and C. F. Mela (2008). Database paper: The iri marketing data set. *Marketing Science 27*(4), 745–748.

Burda, M. and R. Daviet (2023). Hamiltonian sequential monte carlo with application to consumer choice behavior. *Econometric Reviews 42*(1), 54–77.

Dai, C., J. Heng, P. E. Jacob, and N. Whiteley (2022). An invitation to sequential monte carlo samplers. *Journal of the American Statistical Association 117*(539), 1587–1600.

Dawes, J. G. (2018, January). Price promotions: examining the buyer mix and subsequent changes in purchase loyalty. *Journal of Consumer Marketing 35*(4), 366–376. Publisher: Emerald Publishing Limited.

Dew, R. (2024). Adaptive preference measurement with unstructured data. *Management Science 0*, 0.

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

Doucet, A., A. Smith, N. de Freitas, and N. Gordon (2001). *Sequential Monte Carlo Methods in Practice*. Information Science and Statistics. Springer New York.

Drechsler, S., P. S. Leeflang, T. H. Bijmolt, and M. Natter (2017, January). Multi-unit price promotions and their impact on purchase decisions and sales. *European Journal of Marketing 51*(5/6), 1049–1074. Publisher: Emerald Publishing Limited.

Dubé, J.-P. and S. Misra (2023). Personalized pricing and consumer welfare. *Journal of Political Economy 131*(1), 131–189.

Eliaz, K. and R. Spiegler (2011). Consideration sets and competitive marketing. *The Review of Economic Studies 78*(1), 235–262.

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

Gabel, S. and A. Timoshenko (2022). Product choice with large assortments: A scalable deep-learning model. *Management Science 68*(3), 1808–1827.

Goan, E. and C. Fookes (2020). Bayesian neural networks: An introduction and survey. In K. L. Mengersen, P. Pudlo, and C. P. Robert (Eds.), *Case Studies in Applied Bayesian Data Science: CIRM Jean-Morlet Chair, Fall 2018*, pp. 45–87. Springer International Publishing.

Guadagni, P. M. and J. D. Little (1983). A logit model of brand choice calibrated on scanner data. *Marketing Science 2*(3), 203–238.

Haghani, M., M. Bliemer, and D. A. Hensher (2021). The landscape of econometric discrete choice modelling research. *Journal of choice modelling 40*(C), S1755534521000361.

Hornik, K., M. Stinchcombe, and H. White (1989). Multilayer feedforward networks are universal approximators. *Neural Networks 2*(5), 359–366.

Jospin, L. V., H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun (2022). Hands on bayesian neural networks a tutorial for deep learning users. *IEEE Computational Intelligence Magazine 17*(2), 29–48.

Kim, C., A. N. Smith, J. Kim, and G. M. Allenby (2023). Outside good utility and substitution patterns in direct utility models. *Journal of choice modelling 49*, 100447.

Kim, D. S., S. Lee, T. Hur, J. Kim, and G. M. Allenby (2023). A direct utility model for access costs and economies of scope. *Management Science 70*(6), 3381–4165.

Kim, J. (2019, January). The impact of different price promotions on customer retention. *Journal of Retailing and Consumer Services 46*, 95–102.

Lee, S. and G. M. Allenby (2014). Modeling indivisible demand. *Marketing Science 33*(3), 364–381.

Lee, S., J. Kim, and G. M. Allenby (2013). A direct utility model for asymmetric complements. *Marketing Science 32*(3), 454–470.

Leimkuhler, B. and S. Reich (2004). *Simulating Hamiltonian Dynamics*. Cambridge University Press.

Levy, S. (2024). *Essays on Bayesian Machine Learning in Marketing*. Ph. D. thesis, Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, USA.

Liu, H., Y. Ong, X. Shen, and J. Cai (2020, 07). When gaussian process meets big data: A review of scalable gps. *IEEE Transactions On Neural Networks And Learning Systems 31*(11), 4405–4423.

Loken, C., D. Gruner, L. Groer, R. Peltier, N. Bunn, M. Craig, T. Henriques, J. Dempsey, C.-H. Yu, J. Chen, L. J. Dursi, J. Chong, S. Northrup, J. Pinto, N. Knecht, and R. V. Zon (2010). Scinet: Lessons learned from building a power-efficient top-20 system and data centre. *Journal of Physics: Conference Series 256*(1), 012026.

Marshall, P. and E. T. Bradlow (2002). A unified approach to conjoint analysis models. *Journal of the American Statistical Association 97*(459), 674–682.

Mas-Colell, A., M. D. Whinston, and J. R. Green (1995). *Microeconomic theory*. Oxford university press New York.

McCoy, J., R. Ciulli, and E. Bradlow (2022). Two-for-one conjoint: Bayesian cross-category learning for shared-attribute categories. Available at SSRN 4136593.

McFadden, D. and K. Train (2000). Mixed mnl models for discrete response. *Journal of Applied Econometrics 15*(5), 447–470.

Neal, R. M. (1996). Bayesian learning for neural networks. In *Lecture Notes in Statistics*, Volume 118, pp. 1–204. Springer-Verlag.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*. Chapman & Hall / CRC Press.

Ponce, M., R. van Zon, S. Northrup, D. Gruner, J. Chen, F. Ertinaz, A. Fedoseev, L. Groer, F. Mao, B. C. Mundim, M. Nolta, J. Pinto, M. Saldarriaga, V. Slavnic, E. Spence, C.-H. Yu, and W. R. Peltier (2019). Deploying a top-100 supercomputer for large parallel workloads: The niagara supercomputer. In *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (Learning)*, PEARC '19, New York, NY, USA. Association for Computing Machinery.

Radder, L. and W. Huang (2008). High-involvement and low-involvement products: A comparison of brand awareness among students at a south african university. *Journal of Fashion Marketing and Management: An International Journal 12*(2), 232–243.

Roberts, J. H. and J. M. Lattin (1991). Development and testing of a model of consideration set composition. *Journal of Marketing Research 28*(4), 429–440.

Ruiz, F. J., S. Athey, and D. M. Blei (2020). Shopper. *The Annals of Applied Statistics 14*(1), 1–27.

Shaddy, F. and L. Lee (2020, February). Price Promotions Cause Impatience. *Journal of Marketing Research 57*(1), 118–133. Publisher: SAGE Publications Inc.

Steel, M. F. J. (2020, September). Model averaging and its use in economics. *Journal of Economic Literature 58*(3), 644–719.

Toubia, O. (2018). Conjoint analysis. In *Handbook of marketing analytics: Methods and applications in marketing management, public policy, and litigation support*, pp. 52–76. Edward Elgar Publishing Northampton, MA.