Set-wise Attribute Normalization:
A Neural Decision Model for Discrete Choice

Remi Daviet*
Wharton Marketing Department, University of Pennsylvania
November 4, 2019

Abstract

We propose a model seeking to elaborate on the role that choice set composition, a form of context effect, plays in a discrete choice problem through a normalization of the perceived value of each product attribute. Our model extends the comprehension of context effects beyond the classical three products cases of decoy, compromise and similarity. Specifically, we generalize a state-of-the-art class of models stemming from recent research on neural normalization to a multi-attribute choice setting. We impose a structural model based on the neural coding allowing for a particular form of correlation between product utilities. We highlight the properties of the model with an experimental application to credit card choices and a real-world empirical application to the deodorant market. We find evidence for attribute-based normalizing behavior in both settings. Understanding this normalization phenomenon allows firms to optimize their products portfolio with options whose main purpose is to boost the sales of their other products.

Keywords: Context Effect, Discrete Choice, Consumer Neuroscience

1 Introduction

Individual level market data has seen an exponential increase in availability with the explosion of online commerce. Newly available data in turn facilitate the development and testing of new theories on consumer behavior, aided by sophisticated numerical methods of implementation. This paper contributes to the stream of literature analyzing the role of the choice set – the collection

*rdaviet@wharton.upenn.edu
of all choice alternatives available to a decision-maker – plays in the individual decision-making process.

It is well established that the composition of the choice set can influence relative choice probabilities. For instance, a preference relation on the choice set containing two alternatives can be reversed by introducing a third alternative. Recently, machine learning methods, and especially artificial neural networks have allowed us to gain considerable predictive power accounting for various aspects of the entire choice set. Unfortunately, these methods usually provide little insight on the mechanisms underlying the decision process. Machine learning finds patterns in a dataset called the training set and uses these learned patterns to predict choice. These methods show limitations when counterfactuals diverge from the context of the training set used. A more promising approach would rely on a structural model of decision allowing for interpretation of the estimated parameters. The structure would ideally mimic the real human decision process. Advances in neuroscience have allowed us to gain insight on the decision mechanisms involved in choice.

A specific type of neural process, termed normalization, has been found in multiple areas of the brain and across species. Based on insights into this neural process, Webb, Glimcher, and Louie (2016) introduced a structural model called Value Divisive Normalization (VDN) to capture specific types of context effects when product characteristics are not observed. In this paper, we extend this approach to a multi-attribute framework. This extension, Set-wise Attribute Normalization (SAN), can capture multi-attribute context effects such as asymmetric dominance, compromise and similarity. The model has strong implications for companies looking to optimize their product portfolio, as in induces the possibility of creating products whose sole purpose is to boost the sales of another product. While the known context effects have been used in marketing for this purpose, our framework is more general and extends to cases where more than 3 products are offered, which is likely to be the case in most markets.

The structure of this paper is as follows. In Section 2, we briefly review the relevant literature on context effects. In Section 3, we first introduce the SAN model and provide neuroscientific evidence to support our approach (Section 3.1). We then highlight the ability of the SAN model to capture context effects (Section 3.2).

In the second part of the paper, we focus on estimation. We first introduce a noise structure on the utilities and explain how to compute choice probabilities in Section 4. We then impose a Hierarchical Bayesian structure for modeling individual heterogeneity. We also provide details about the Bayesian estimation method used. In particular, we propose a Sequential Monte Carlo
(SMC) approach allowing for real-time estimation and updating of the preference parameters. In Section 5 we apply our model to an experimental dataset of credit card choices allowing to test the model in a controlled environment. Finally, in Section 6 we use an dataset of deodorant purchase to detect normalization in an empirical context. We conclude the paper with an overview of possible model extensions and potential future research. Some proofs and complements are provided in Appendices.

2 Discrete Choice and Context Effects

We want to study the choice of an individual presented with a choice set of alternatives $J = \{1,...,J\}$. Each alternative $j$ can be represented by its attributes $x_j = (x_{jk}; k = 1,...,K)$. The context is the collection of attributes in the choice set: $X = (x_j; j \in J)$. Each alternative has the same number of attributes, with each restricted to be non-negative. The attributes may be quantitative (price, size,...), or qualitative (brand, color,...). Qualitative attributes are represented by the sum of the perceived values of the qualities they possess. Such qualities are usually represented by a vector of dummy variables $(x_{jkl} \in \{0,1\}; l = 1,...,L_k)$, where attribute $k$ has $L_k$ possible qualities and $x_{jkl} = 1$ if product $j$ possesses that quality (see Section 4.3.1). The individual chooses a unique alternative in the choice set $J$.

It is well understood that the composition of a choice set can influence preferences. Several examples of such influences are grouped under the label of context effects. The first type of context effect is the result of changing the choice set size, independently of the characteristics of the newly added alternatives. We are however interested in the type of context effects where these characteristics are relevant. Most studies eliciting these effects compare the choice proportions between a binary case where two alternatives are offered, and a trinary case where an additional option is added. In these examples, the choice option reinforced by the addition of a third option is usually called the target, and will be denoted $T$ with characteristics $x_T$, while the other alternative is called the competitor and will be denoted $A$ with characteristics $x_A$. We are interested in explaining the probability $P_J(T)$ of the target $T$ being chosen given a specific choice set $J$. We sometime also discuss the conditional probability given that either $T$ or $A$ has been chosen: $P_J(T|\{T,A\})$ Note that context effects depend on attribute characteristics and individual preferences. It is thus expected that different individuals show different choice patterns that may or may not agree with the context effects described below. Our model allows to capture and explain these variations in
context effects under a unifying framework.

2.1 Asymmetric Dominance

For the attraction, or decoy, or asymmetric dominance effect, consumers are more likely to choose a target product \( T \) than its competitor \( A \) if the target is presented along with a decoy product \( D \) dominated by the target. By dominated, we mean that \( T \) is more desirable than \( D \) on every attribute dimension: \( x_{Tk} > x_{Dk}, \forall k \). In choice experiments where an alternative is dominated, the probability of the dominated alternative being chosen is typically close to 0, as would be expected.

The asymmetric dominance effect occurs when the addition of a dominated alternative shifts relative choice probabilities toward the alternative that dominates it and can generate a violation.

Figure 1: An illustration of: the dominance case (top-left), the asymmetric dominance case (top-right), the compromise case (bottom-left) and the similarity case (bottom-right).
of regularity (Huber, Payne, and Puto, 1982):

\[
\frac{P_{\{x_T, x_A, x_D\}}(T|\{T, A\})}{P_{\{x_T, x_A, x_D\}}(A|\{T, A\})} > \frac{P_{\{x_T, x_A\}}(T)}{P_{\{x_T, x_A\}}(A)}.
\]

When it exists, this effect has important consequences from a marketing point of view: the decoy’s purpose is to increase the sales of the target, to be sold. Note that the practical relevance has been subject to a recurrent debate (see Crosetto and Gaudeul, 2016; Frederick, Lee, and Baskin, 2014; Huber, Payne, and Puto, 2014; Simonson, 1989, 2014; Yang and Lynn, 2014).

2.2 Compromise Effect

The compromise effect is closely related to the asymmetric dominance effect and occurs when the introduction of an extreme alternative \(x_C\) shifts the relative choice probabilities toward the alternative that holds the intermediate values on each attribute dimension (Simonson, 1989; Simonson and Tversky, 1992). An extreme alternative has a very high attribute value in one dimension \((x_C > x_T > x_A)\), and very low in another dimension \((x_C < x_T < x_A)\). The asymmetric dominance effect and the compromise effect are often studied together in many applications.

2.3 Similarity Effect

Finally, the similarity or substitution effect occurs when an added choice alternative decreases the choice probability of similar alternatives more than dissimilar ones, potentially generating a preference reversal (Tversky, 1972). The concept of similarity also depends on the definition of distance between alternatives. This effect is not always observed in practice. Indeed, the reverse effect can be elicited by merely changing the presentation format of questions in an experiment (Cataldo and Cohen, 2018; Noguchi and Stewart, 2014).

In the following section, we build a model explaining these empirical effects using functions inspired by the neuroscience literature.

3 Set-wise Attribute Normalization

In this section, we first describe the neural foundations for Set-wise Attribute Normalization (SAN). We then study model properties and compare them to the empirical context effects from Section 2.
3.1 Model and Neural Foundations

Figure 2: Diagram representing the value normalization process for alternative 1’s $k$-th attribute. The observed value $x_{1k}$ is normalized relative to the average of the other alternatives’ $k$-th attributes. The output of this process is the normalized value $V_k(1; X)$.

Figure 3: Diagram representing how the alternative $j$’s subjective value $U(j; X)$ is computed from its attributes’ normalized values.

Our model has three layers. Attribute magnitudes are first observed by the individual and then normalized with respect to context and beliefs (Figure 2). The normalized value is then sent through a synapse, that is, a connection between two neurons, to an integrator neuron proceeding to a weighted sum. The result of this sum is the alternative’s subjective value (Figure 3). Finally the individual chooses the alternative with the highest subjective value.
In the following subsections, we discuss the various computations taking place in our neural model. We first introduce the normalization process, and then discuss the aggregation.

3.1.1 Divisive Normalization

Figure 4: Naka-Rushton function with $\mu = 1$ and various values for the slope parameter $r$ (top-left) and Naka-Rushton function with $r = 2$ and various values for $\mu$ (bottom-left). In these plots, the x-axis scale is logarithmic.

The Divisive Normalization (DN) principle is a representation of a well known neural mechanism that we discuss below. Neuroscience has identified a set of canonical neural computations that the brain relies on. These computations are the results of a neuron or group of neurons and can be found across brain regions and functions. The main canonical neural computations are exponentiation, (thresholding), linear filtering, (weighted averaging), and divisive normalization (DN) (see Carandini and Heeger, 2011, for a review).

**Definition 1.** A valuation function follows the DN principle if it can be written as:

$$
\nu(x_{jk}, \mu_k) = \frac{x_{jk}^r}{\mu_k^r + x_{jk}^r} = \frac{1}{1 + e^{-r_k \log(x_{jk}) - \log(\mu_k)}},
$$

where $\mu_k > 0$ and $r_k > 0$ are two parameters, and the perceived attribute $x_{jk}$ is non negative.

DN has been shown to be optimal given neurobiological constraints (Gottschalk, 2002; Schwartz and Simoncelli, 2001; Valero and Navarro, 2003). Encoding information is costly for the brain and done under a limited amount of resources. Neural activity involves chemical and thermodynamic processes making it appear stochastic. The effect of this stochastic noise can be attenuated by increasing the number of neurons encoding the information. However, increasing the number of
neurons also increases the cost for the brain. In addition, the neural signal is bounded, which imposes some type of relative coding. Given these constraints, DN maximizes information and minimizes errors.

In our context, \( v(x_{jk}, \mu_k) \) is interpreted as the subjective value associated with a perceived attribute magnitude of \( x_{jk} \). This equation is also known as the Naka-Rushton function and has been observed in neurons in various areas of the brain and across species by electro-physiological measurements of the signal (Kleinschmidt and Dowling, 1975; Naka and Rushton, 1966). When plotted on a log scale, this function has a sigmoid shape with inflexion point at \( \mu_k \) and slope controlled by \( r_k \). When written in log scale, it becomes the standard logit transformation. We show in Figure 4 a comparison of Naka-Rushton functions with various parameter values. In the rest of this paper, we refer to \( \mu_k \) as the reference point, as it is the main interpretation of its role in our context. While DN has originally been studied in the perception literature, it has been shown to be a fundamental computation involved in decision with both neural and behavioral evidence (Holper et al., 2017; LoFaro et al., 2014; Louie, Grattan, and Glimcher, 2011).

The valuation formula (1) assumes that \( x_{jk} > 0 \) in order to provide a meaning to a fractional power \( x^r \). Thus, if the initial attribute characteristic is negative, the brain will first transform it into a perceived input, rendering it positive. Such implicit transformation will also be done to treat different possible units of the initial \( x \), making comparable for instance an information in miles and in meters. The distinction between the two notions is elaborated in Section 4.3.1, where we discuss the econometric model and identification issues.

### 3.1.2 Value Divisive Normalization

In the general DN formula, the value is not a function of the context. However, there is mounting evidence that in a valuation context the reference point is a type of average of a choice set’s perceived values. Webb, Glimcher, and Louie, 2016 implemented this principle with the Value Divisive Normalization (VDN) where the reference point is computed with a generalized mean formula. We consider below a slightly different form, due to Carandini and Heeger, 2011, and integrate the VDN’s generalized mean into its reference point:
Definition 2. The Value Divisive Normalization of an attribute \( k \) in a context \( X \) is:

\[
V_k(j; X) = \frac{x_{jk}^{r_k}}{((1 - w_k)\mu_{k0} + w_k(\frac{1}{J} \sum_{j'} x_{j'k}^c)^{1/c})r_k + x_{jk}^{r_k}}
\]

\[
= \frac{\mu_k(X)^{r_k} + x_{jk}^{r_k}}{1 + e^{-r_k(\log(x_{jk}) - \log(\mu_k(X)))}}
\]

(2)

with \( \mu_k(X) = (1 - w_k)\mu_{k0} + w_k(\frac{1}{J} \sum_{j'} x_{j'k}^c)^{1/c}, 0 \leq w_k \leq 1 \) and \( c > 0 \).

Mathematically, VDN is nested in SAN with each product being represented by a unique categorical attribute and \( r_k = 1 \). The reference point can be interpreted as a weighted average between a pooled response from all receptor neurons and prior beliefs represented by \( \mu_{k0} \). This allows us to interpret \( w_k \) as the weight of the context in our beliefs. The generalized mean has different interpretation. It is the geometric mean when \( c \) tends to 0, the arithmetic mean when \( c = 1 \), a quadratic mean when \( c = 2 \), and is equal to the maximum of the \( x_{jk} \), \( j = 1, ..., J \), when \( c \) tends to infinity.

This form provides us with another interpretation of the valuation mechanism. The field of psychophysics has documented for over a century that we perceive magnitudes as a log function of their actual value, a fact known as the Weber-Fechner law. The law can be explained by our specification as the normalized value is a function of the log of the attribute magnitude compared to the log of the perceived mean. On the other hand, neuroscience has accumulated evidence that in the context of valuation, subjective values are normalized with respect to the distribution of potentially available rewards (Rigoli, Friston, and Dolan, 2016). Since our functional form in log scale is the cdf of a logistic distribution, the value of an attribute can be interpreted as the log magnitude normalized with respect to a logistic distribution of mean \( \ln(\mu_k(X)) \) and scale \( 1/r \). This provides interpretation for the term valuation: the function indicates if an attribute characteristic is low or high relative to a reference point and a distribution.

3.1.3 Attribute Values Aggregation

Once normalized attributes values are computed, they are aggregated into an overall option value. The average overall value of a choice option \( \tilde{U}(j; X) \) is the result of a weighted sum, mimicking the role of integrator neurons:

\[
\tilde{U}(j; X) = \sum_k^K \beta_k V_k(j; X),
\]
where the coefficient $\beta_k$ are either positive or negative.

The neurons in charge of valuation are connected to a second group of neurons integrating the various valuation signals. In order to understand how the aggregation is done, we need to describe how neurons communicate with each other. Each neuron sends a signal in the form of an electrical impulse through an extension called an axon. At their ends, axons transmit the signal to the next neuron through a connection called a synapse. An impulse (action potential) reaching the end of an axon results in an average change of voltage in the receiving (post-synaptic) neuron. These resulting changes can be positive or negative and have different magnitudes depending on which synapse the signal came in. The degree of average change resulting from an impulse coming from a particular synapse is known as synaptic strength. This strength is known to evolve over time in a phenomenon known as synaptic plasticity. In our model, $\beta_k$ is the average synaptic strength between a group of valuation neuron $k$ and integrator neurons. The synaptic strength can also be interpreted as a representation of a taste for product characteristics (see below).

3.1.4 Set-wise Attribute Normalization (SAN)

Synapses experience short term variations lasting at most a few minutes (Zucker and Regehr, 2002) and long term variations known as long term potentiation and depression (Frey and Morris, 1997; Malenka and Nicoll, 1999). The synaptic strength can be altered by a large panel of factors such as environment, emotion, or experience. We expect the synaptic strength to be changing between each decision and we have to take this effect into account in our modeling by adding a random term $\epsilon_{kt}$ in the synaptic strength. Moreover, there might be unobserved factors affecting the overall value of a choice option; we represent these factors through an additive term $\xi_{jt}$.

**Definition 3.** The Set-wise Attribute Normalization model is defined through the subjective values:

$$U_t(j; X) = \sum_{k=1}^{K} (\beta_k + \epsilon_{kt})V_k(j; X) + \xi_{jt},$$

(3)

where $\epsilon_{kt}, \xi_{jt}$ are stochastic shocks.

The individual then chooses the alternative with the highest subjective value. There are two sources of variation here: i) variation in the tastes for product characteristics $\epsilon_{kt}$ and ii) the unobservable factors $\xi_{jt}$. Both sources have to be taken into account to evaluate choice probabilities.

We assume that both $\epsilon_{kt}$ and $\xi_{jt}$ are normally distributed with zero mean. We also assume independence in $(k,t)$ for $\epsilon_{kt}$, and in $(j,t)$ for $\xi_{jt}$. The independence assumption is kept for simplicity.
and convenience, but can readily be relaxed. This noise structure falls into the probit family of discrete choice models studied by Hausman and Wise, 1978. We discuss the implications in Section 4.

### 3.2 Context Effects

In this section we analyze the properties of the SAN model. After illustrating how the function behaves in the simple one attribute case, we extend the analysis to multi-attribute context effects. In this section we set attribute weights equal to 1 and assume a noise free environment:

$$ U(j; X) = \sum_{k}^{K} V_k(j; X). $$

#### 3.2.1 Influence of the Reference Point

![Figure 5: $v(6, \mu_k) - v(2, \mu_k)$ as a function of the reference point $\mu_k$ with various values for $r$.](image)

To understand how our model constrain context effects, we first have to focus on the influence of the reference point $\mu_k$ on the normalization function:

$$ v(x_{jk}, \mu_k) = \frac{x_{jk}^{r_k}}{\mu_k^{r_k} + x_{jk}^{r_k}}. $$

More precisely, we are interested in the behavior of $\Delta v(x_{1k}, x_{2k}, \mu_k) = v(x_{2k}, \mu_k) - v(x_{1k}, \mu_k)$ as $\mu_k$ varies. Without loss of generality, we consider that $x_{2k} > x_{1k}$ and thus $\Delta v(x_{1k}, x_{2k}, \mu_k) > 0, \forall \mu_k > 0$. The derivative of $\Delta v(x_{1k}, x_{2k}, \mu_k)$ with respect to the reference point is:

$$ \frac{\partial \Delta v(x_{1k}, x_{2k}, \mu_k)}{\partial \mu_k} = r \mu_k^{r-1} \left( \frac{x_{1k}^{r_k}}{(x_{1k}^{r_k} + \mu_k^{r_k})^2} - \frac{x_{2k}^{r_k}}{(x_{2k}^{r_k} + \mu_k^{r_k})^2} \right). $$
The derivative admits 3 roots: at 0, $\infty$ and $\sqrt[3]{x_{1k}x_{2k}}$ (see Appendix A.1). The distance between the normalized values of $x_{1k}$ and $x_{2k}$ is maximized when $\mu_k^* = \sqrt[3]{x_{1k}x_{2k}}$, or equivalently when $\ln(\mu_k^*) = \frac{\ln(x_{1k}) + \ln(x_{2k})}{2}$. The difference between the normalized attribute values decreases as the distance between $\mu_k$ and $\mu_k^*$ increases. Different values of $r$ only change the magnitude, but not the behavior of the function.

Another important result is that there are always 2 values for $\mu_k$ yielding the same distance $\Delta v(x_{1k}, x_{2k}, \mu_k)$ except for the maximum. This result will be used in subsequent sections. Let us denote $\mu_k^{(1)}$ and $\mu_k^{(2)}$ the solutions to the equation $\Delta v(x_{1k}, x_{2k}, \mu_k) = b$, where $b$ is an arbitrary number. If they exists, these solutions satisfy the relation (See Appendix A.2):

$$\mu_k^{(1)} = \frac{x_{1k}x_{2k}}{\mu_k^{(2)}}.$$  \hspace{1cm} (4)

To better illustrate the value function behavior, we plot $\Delta v(x_{1k}, x_{2k}, \mu_k)$ for $x_{1k} = 2$, $x_{2k} = 6$ and for different values of $\mu_k$ and $r_k$ (see Figure 5). This function admits a global maximum at 3.46, and higher values for $r$ increase the magnitude of the maximum difference.

### 3.2.2 The Effect of a Third Alternative

Let us now allow $\mu_k$ to depend on the context and study the change in values of two alternatives when a third alternative is added to the choice set. We here work with a simplified model where the weight of the context is maximal $w_k = 1$. The value function in our case simplifies to:

$$V_k(j; X) = \frac{x_{jk}^{r_j}}{(\frac{1}{2} \sum_{j'=1}^{J} x_{jk}^{r_j})^{r/c} + x_{jk}^{r_j}}.$$  

In the case of three alternatives, the behavior of $\Delta v(x_{1k}, x_{2k}, \mu_k(X))$ with respect to $x_{3k}$ is similar to what was described in the previous section concerning $\mu_k$. The function’s derivative is:

$$\frac{\partial \Delta v(x_{1k}, x_{2k}, \mu_k(X))}{\partial x_{3k}} = \frac{\partial \mu_k(X)}{\partial x_{3k}} \left[ \frac{\partial \Delta v(x_{1k}, x_{2k}, \mu_k)}{\partial \mu_k} \right]_{\mu_k = \mu_k(X)}$$

$$= r \mu_k(X)^{-c} \left( \frac{x_{1k}^{r_k}}{(x_{1k}^{r_k} + \mu_k(X)^{r_k})^2} - \frac{x_{2k}^{r_k}}{(x_{2k}^{r_k} + \mu_k(X)^{r_k})^2} \right) x_{3k}^{c-1}.$$  

$\Delta v(x_{1k}, x_{2k}, \mu_k(X))$ admits a maximum at $x_{3k}^* = (3\sqrt[3]{x_{1k}x_{2k}} - x_{1k}^c - x_{2k}^c)^{1/c}$, if $x_{3k}^* > 0$; otherwise the maximizer is 0. This maximizer is strictly positive for $\frac{x_{2k}}{x_{1k}} < (\frac{3 + \sqrt{3}}{2})^{2/c}$ (approximately $6.854^{1/c}$) (see Appendix A.3). We can note that $x_{3k}^*$ is also the solution for $(\frac{1}{3} \sum_{j'=1}^{J} x_{j'k}^{c})^{1/c} = \sqrt[3]{x_{1k}x_{2k}}$, which maximizes $\Delta v(x_{1k}, x_{2k}, \mu_k)$ over all possible values for $\mu_k$. Consequently, the maximum value for $\Delta v(x_{1k}, x_{2k}, \mu_k)$ with three alternatives is always at least as high as $\Delta v(x_{1k}, x_{2k}, \mu_k)$ with two alternatives.
Figure 6: $V_k(j; X)$ as a function of $x_{3k}$ for $x_{1k} = 2$, $x_{2k} = 6$ and $r = 1$, with $c = 0.5$ (top-left), $c = 1$ (top-right) and $c = 2$ (bottom-left). Bottom-right: $\Delta v(x_{1k},x_{2k},\mu_k)$ as a function of $x_{3k}$ for $r = 1$ with various values for $c$. Dashed lines represent the binary values or value differences.

We also want to know for which values of $x_{3k}$ the distance between the normalized values of $x_{1k}$ and $x_{2k}$ is greater than in the binary case. We find two solutions presented in Proposition 1.

**Proposition 1.** For $x_{2k} > x_{1k}$, $X_2 = \{x_{1k}, x_{2k}\}$, $X_3 = \{x_{1k}, x_{2k}, x_{3k}\}$:

1. $x_{3k}^* = \arg\max_{x_{3k}} \Delta v(x_{1k},x_{2k},\mu_k(X_3)) = \begin{cases} (3\sqrt{x_{1k}x_{2k}}^c - x_{1k}^c - x_{2k}^c)^{1/c}, & \text{if } x_{2k}^c - x_{1k}^c \leq \left(\frac{3+\sqrt{3}}{2}\right)^{2/c}, \\ 0, & \text{otherwise}. \end{cases}$

2. $x_{3k}^* < x_{2k}$ (see Appendix A.4 for a proof)

3. $\Delta v(x_{1k},x_{2k},\mu_k(X_3)) \geq \Delta v(x_{1k},x_{2k},\mu_k(X_2))$ for

$$x_{3k} \in \begin{cases} \left[\left(\frac{6x_{1k}^c x_{2k}^c}{x_{1k}^c + x_{2k}^c} - x_{1k}^c - x_{2k}^c\right)^{1/c}, \left(\frac{x_{1k}^c + x_{2k}^c}{2}\right)^{1/c}\right], & \text{if } \frac{x_{2k}^c}{x_{1k}^c} < (2 + \sqrt{3})^{1/c}, \\ 0, \left(\frac{x_{1k}^c + x_{2k}^c}{2}\right)^{1/c}, & \text{otherwise}. \end{cases}$$
We present in Figure 6 an example where $x_{1k} = 2$ and $x_{2k} = 6$ to illustrate the behavior of the function graphically. We note that this function reaches one maximum at the expected $x_{3k}^*$ value (e.g. $x_{3k}^* = 2.39$ for $c = 1$).

### 3.2.3 Asymmetric Dominance

![Attribute Magnitudes](image1)

![Attribute Value Shift](image2)

![Options Value Shift](image3)

Figure 7: Illustration of an asymmetric dominance effect for $c = 1$ and $r = 5$. Left: Attribute magnitudes for the target ($x_T$), the alternative ($x_A$) and the decoy ($x_D$). Center: corresponding normalized values before adding a decoy (blue) and after (red). Right: Overall value for each alternative before adding a decoy (blue) and after (red).

We seek to assess whether our model is able to capture an asymmetric dominance effect. Consider a binary choice set $X_B$ with alternatives $T$ and $A$ such that, $x_{T1} < x_{A1}$, $x_{T2} > x_{A2}$, and $U(T; X_B) = U(A; X_B)$. We want to know if there a trinary choice set $X_T = \{x_T, x_A, x_D\}$ such that $U(T; X_T) > U(A; X_T)$. The decoy region is defined as the set of points satisfying asymmetric dominance: $x_{D1} < x_{T1}$ and $x_{A2} < x_{D2} < x_{T2}$.

By definition, asymmetric dominance effect happens when:

$$\frac{x_{T1}^{r_k}}{(x_{T1}^2 + x_{A1}^2 + x_{D1}^2)^{r/c}} + \frac{x_{T2}^{r_k}}{(x_{T2}^2 + x_{A2}^2 + x_{D2}^2)^{r/c}} > \frac{x_{A1}^{r_k}}{(x_{T1}^2 + x_{A1}^2 + x_{D1}^2)^{r/c}} + \frac{x_{A2}^{r_k}}{(x_{T2}^2 + x_{A2}^2 + x_{D2}^2)^{r/c}}$$

The effect can be driven by two factors: (1) the distance between the normalized values on the first dimension shrinks when $x_D$ is added, and (2) the distance between the normalized values on
the second dimension increases when $x_D$ is added. It is sufficient for one of these factors to be satisfied as long as the resulting effect is stronger than any opposing effect.

On the first dimension, we argue that there exists a point $x_{D1} < x_T$ that satisfies:

$$\Delta v(x_T, x_A, \mu_1(X_T)) < \Delta v(x_T, x_A, \mu_1(X_B)),$$

only if $x_{A1} / x_{D1} < (2 + \sqrt{3})^{1/c}$. This can be shown by considering the cases in which the extreme point $x_{D1} = 0$ shrinks the distance between the normalized values once added to the choice set (see Appendix A.5). In other words, it diminishes the difference in normalized values on the dimension where $x_T$ is worse than $x_A$. The shrinking effect diminishes as $x_{D1}$ increases and can even be reversed.

On the second dimension, we aim to increase the distance between $x_A$ and $x_T$ to make the target appear even more dominant. We already solved that problem in Section 3.2.2 and showed that $x_{D2}^* = \left(3\sqrt{x_T x_A} - x_T^2 - x_A^2\right)^{1/c}$ is optimal if it falls in the decoy region.

Under the conditions mentioned above for the decoy, the model predicts an attraction effect. We provide a graphical example in Figure 7.

3.2.4 Compromise effect

Consider again a binary choice set $X_B$ with target $T$ and an alternative $A$ such that $x_{T1} < x_{A1}$, $x_{T2} > x_{A2}$, and $U(T; X_B) = U(A; X_B)$. We want to know if there is a trinary choice set $X_T = \{x_T, x_A, x_C\}$, with $x_{C1} << x_{T1}$ and $x_{T2} < x_{C2}$ such that $U(T; X_T) > U(A; X_T)$.

In our model, the mechanism generating the compromise effect is very similar to the one behind the asymmetric dominance effect. In fact, it can be argued that they are the same effect. As for asymmetric dominance, the extremely low value of $x_{C1}$ on the first attribute dimension potentially reduces the perceived value difference between $x_{T1}$ and $x_{A1}$. As the value $x_{C2}$ is larger than $x_{T2}$, the normalized value difference between the target and the alternative diminishes as $x_{C2}$ increases (see Section 3.2.2 and Appendix A.4). Consequently, the compromise effect is maximal when $x_{C1} = 0$ and $x_{C2} = x_{T2}$. The model also implies that the strongest compromise effect is weaker than the strongest decoy effect. We provide a graphical example in Figure 8.

3.2.5 Similarity effect

Let us consider again a binary choice set $X_B$ with a target $T$ and an alternative $A$ such that $x_{T1} < x_{A1}$, $x_{T2} > x_{A2}$, and $U(T; X_B) = U(A; X_B)$. We want to know if there is a trinary
Figure 8: Illustration of a compromise effect for $c = 1$ and $r = 5$. Left: Attribute magnitudes for the target ($x_T$), the alternative ($x_A$) and the added alternative ($x_C$). Center: corresponding normalized values before adding the third alternative (blue) and after (red). Right: Overall value for each alternative before adding the third alternative (blue) and after (red).

choice set $X_T = \{x_T, x_A, x_S\}$, with the additional option $S$ in a region close to $A$, such that $U(T; X_T) > U(A; X_T)$.

We will consider the case where $x_S = x_A$. Adding this third alternative to the choice set has two effects. First, it reduces the difference in normalized values on the first dimension, where $x_A$ dominates (see Appendix A.7). Second, it increases the difference in normalized values on the second attribute dimension if $\frac{x_T}{x_A} < 2^{1/c}$, and decreases it if $\frac{x_T}{x_A} > 2^{1/c}$. In the first case, we see a similarity effect as on both dimensions $x_T$ appears more desirable than in the binary case. However, if $\frac{x_T}{x_A} < 2^{1/c}$, the presence of a similarity effect will depend on the magnitude of the effect on each dimension. Since the value functions are continuous in $x_S$, a similar effect is to be expected in the region around $x_A$. We provide a graphical example in Figure 9.
Figure 9: Illustration of a similarity effect for $c = 1$ and $r = 5$. Left: Attribute magnitudes for the target ($x_T$), the alternative ($x_A$) and the added alternative ($x_S$). Center: corresponding normalized values before adding the third alternative (blue) and after (red). Right: Overall value for each alternative before adding the third alternative (blue) and after (red).
3.2.6 Summary of Context Effects with Three Alternatives

Figure 10: The colors represent the value of $U(T; X_T) - U(A; X_T)$ as a function of $x_3$ for $r = 2$ when $X_T = \{x_T, x_A, x_3\}$. The blue line represents $x_T$'s indifference set for $x_T$ in the binary case (see Section 3.2.6).

We summarize the potential context effects in Figure 10, where we specify a target $T$ and an alternative $A$ which are situated at the same indifference curve in the binary case (the indifference set for $T$ is represented by a blue line). The colored areas represent the difference in overall value between $T$ and $A$ when a third alternative is added at a point $x_3$ in the attribute space. We can see the continuity of the effect between the decoy region and the compromise region. We can also observe the compromise and asymmetric dominance effect progressively weakening and being replaced by the similarity effect as $x_3$ gets closer to $x_T$.

The analysis performed in Section 3.2 provides some insight into the flexibility of the functional form retained for the valuation function $V$. However, it does not take into account the aggregation weights $\beta_k$, and the fact that the utilities will become random. This latter aspect is important since with random utilities, two types of preferences can be defined: i) the unobserved preference ordering based on the utilities themselves, and ii) the observed preference ordering based on the choice probabilities. In the random utility framework, the second preference ordering matters. Note also that the illustrations have been given with the target and the competitor on the same
indifference curve in the binary case. It would be useful to explore the strength of the context effects when the competitor is positioned on higher or lower indifference curves (see Crosetto and Gaudeul, 2016).

4 Econometric Model

In this section we first define the model, and explain how to compute the choice probabilities and the corresponding likelihood function. We discuss identification of the individual level parameters. Then, to account for individual heterogeneity, we introduce a hierarchical Bayesian structure, with intermediate hyperparameters. We detail the choice of prior distributions for both the parameters and hyper parameters. Finally we propose an estimation procedure based on Sequential Monte Carlo methods.

4.1 The Parametric Model

We observe a series of individual choices \( y_{it} \in J_{it} \). The choice sets are represented by \((J_{it}, X_{it})\), where \( J_{it} \subset \{1, \ldots, J\} \) and \( X_{it} \) provides the attribute levels for each of the alternatives in \( J_{it} \). The choice sets vary across individuals and time, that is \( i = 1, \ldots, n \), and \( t = 1, \ldots, T_i \). The panel data set is assumed unbalanced and the total number of observations is \( N = \sum_{i=1}^{N} T_i \). We seek to estimate the individual decision model from the observations \( y_{it} \) defined by:

\[
y_{it} = \arg \max_{j \in J_{it}} U_{it}(j; X_{it}), \text{ where}
\]

\[
U_{it}(j; X_{it}) = \sum_{k=1}^{K} (\beta_{ik} + \epsilon_{ikt}) V_{ik}(j; X_{it}) + \xi_{ijt},
\]

\[
V_{ik}(j; X_{it}) = \frac{x_{jk}^{r_{ik}}}{\mu_{ik}(X)^{r_{ik}} + x_{jk}^{r_{ik}}},
\]

\[
\mu_{ik}(X_{it}) = (1 - w_{ik})\mu_{ik0} + w_{ik}\left(\frac{1}{\text{card}(J_{it})} \sum_{j \in J_{it}} x_{jk}^{c_i}\right)^{1/c_i},
\]

where \( x_{jk} \) is a function of observed explanatory variables (see the discussion in Section 4.3.1). We collect individual parameters in \( \theta_i \) as follow:

\[
\theta_i = (a_{ik}, \beta_{ik}, r_{ik}, \mu_{0ik}, w_{ik}, c_i, \sigma_{ik}; k = 1, \ldots, K), i = 1, \ldots, n,
\]

and denote \( \theta = (\theta_1, \ldots, \theta_n) \).
The total number of parameters is relatively large, equal to \( \dim(\theta) = n(1 + 5K + \sum_{k=1}^{K} L_k) \). It is potentially larger than the number of observations. This curse of dimensionality problem will be solved below by parametrizing heterogeneity and prior distributions.

4.2 Choice Probabilities and Likelihood Function

4.2.1 Choice Probabilities

In this section we first study the choice probabilities for a given individual at a given time and omit the indexes \( i, t \) for clarity. An individual’s subjective value can be rewritten as a sum of a deterministic component and a stochastic component:

\[
U(j; X) = \sum_{k=1}^{K} \beta_k V_k(j; X) + \sum_{k=1}^{K} \epsilon_k V_k(j; X) + \xi_j
\]

with \( \epsilon^*_j(X) = \sum_{k=1}^{K} \epsilon_k V_k(j; X) + \xi_j \). Following Hausman and Wise, 1978, we assume Gaussian distributions:

\[
\epsilon_k \overset{\text{i.i.d.}}{\sim} N(0, \sigma_k)
\]

\[
\xi_j \overset{\text{i.i.d.}}{\sim} N(0, 1),
\]

with the unit variance set to normalize subjective values for scale end ensure identification (see Section 4.3). Consequently, \( \epsilon^*_j \) also follows a Gaussian distribution with a specific covariance structure. Therefore, conditional on \( X \), the stochastic utilities are Gaussian, with mean and conditional variance-covariance matrix:

\[
E(U|X) = \left( \sum_{k=1}^{K} \beta_k V_k(1; X), ..., \sum_{k=1}^{K} \beta_k V_k(J; X) \right)'
\]

\[
V(U|X) = \Sigma(X),
\]

where \( [\Sigma(X)]_{jj'} = \sum_{k=1}^{K} \sigma^2_k V_k(j) V_k(j') + 1 \) for \( j, j' = 1, ..., J \).

By assuming random taste, we introduced another context effect through the second order moments \( \left( \sum_{k=1}^{K} \sigma^2_k V_k(j) V_k(j') + 1 \right) \). Therefore, our model contains two types of context effects. The first one is the result of the characteristic summary \( \left( \sum_{j=1}^{J} x_{jk}^c \right)^{1/c} \) and depends on the parameters \( c \) and \( w_k, k = 1, ..., K \). The second one stems from randomness and depends on \( \sigma_k, k = 1, ..., K \).
Note that to avoid a curse of dimensionality we have assumed independent $\epsilon_k$ (see Hausman and Wise, 1978, for a discussion of this assumption).

The choice probabilities depend on the individual parameters $\theta = (a_k, \beta_k, r_k, \mu_0, w_k, c, \sigma_k; k = 1, \ldots, K)$ and are given by the multi-dimensional integral:

$$P_X(j|\theta) = \int \mathbb{1}\{U(j; X) > U(j'; X); j' \in X, j' \neq j\} \ p(\epsilon^*_1, \ldots, \epsilon^*_J|X) \ d\epsilon^*_1 \ldots d\epsilon^*_J,$$

where $p(\epsilon^*_1, \ldots, \epsilon^*_J|X)$ denotes the probability density for a multivariate normal distribution with mean 0 and variance-covariance $\Sigma(X)$. These probabilities can be numerically approximated with a smoothed accept-reject simulator (e.g. Train, 2009, p.120).

### 4.2.2 Likelihood Function

The corresponding likelihood function for observed choices $y = (y_{it}, i = 1, \ldots, n, t = 1, \ldots, T_i)$ is:

$$l^*(y|\theta, X) = \prod_{i=1}^{n} \prod_{t=1}^{T_i} P_{X_{it}}(y_{it}|\theta), \quad (5)$$

where $X = (X_{it}, i = 1, \ldots, n, t = 1, \ldots, T_i)$ and $\theta = (\theta_i, i = 1, \ldots, n)$.

### 4.3 Identification

#### 4.3.1 Observed Characteristics and Perceived Attribute

Let us now discuss in greater detail the difference between the observed characteristic of an attribute, say $z$, and the perceived attribute $x$.

If $z$ is a polytomous qualitative variable, which can be ordered or not, the variable $z$ can be equivalently represented by a set of dummy variables defined by:

$$x_l = \begin{cases} 
1 & \text{if } z = l \\
0 & \text{otherwise.}
\end{cases}$$

Then the perceived attribute:

$$x = \sum_{l=1}^{L} a_l x_l$$

assigns a value $a_l$ to the qualitative level $l$.

If $z$ is a quantitative real variable, two approaches can be followed. In the first approach, this variable $z$ is transformed into a qualitative variable by segmentation. More precisely we can
introduce a grid of values $\alpha_1 < ... < \alpha_{L-1}$ and define:

$$x_l = \begin{cases} 
1, & \text{if } \alpha_{l-1} < z < \alpha_l \\
0, & \text{otherwise.}
\end{cases}$$

with the convention $\alpha_0 = -\infty$, $\alpha_L = \infty$. Loosely speaking, the perceived attribute is a piecewise transformation of $z$, say $x = h(z)$, that approximates any nonlinear transformation of $z$, when the grid is sufficiently thin. This approach is the one followed in the standard data mining literature.

The second approach often followed in the literature constrains ex-ante the form of this function and assumes $x = z$, which is equivalent to $L = 1$, $a_l = 1$.

Let us now discuss the identification issue. The basic expression of the normalized value $V$ becomes:

$$V = \frac{x^r}{\mu^r + x^r} = \frac{\left(\sum_l a_l x_l\right)^r}{\mu^r + \left(\sum_l a_l x_l\right)^r}.$$ 

We see immediately that the $a_l$ and the $\mu$ are not identifiable since the same value of $V$ is obtained when they are multiplied by a same scalar. In other words, we cannot disentangle the perception step from the valuation step. In order to solve this identification issue, we adopt the following approach: if $x$ is quantitative, $x = z$ and $\mu$ is identifiable. Otherwise, for a qualitative attribute, we assume that $\sum_{l=1}^{L} a_l = 1$. Note that the neuroscience literature has shown that the perceived value for a qualitative variable with two values, say 1 and 2, is different if these values are represented by numbers (1,2), written letters (one, two), or given graphically. Our modeling does not account for that fact.

### 4.3.2 Choice Set and Identification

Another identification issue comes from the variability of the choice sets $J_{it}$. To illustrate the identification issue here, we can consider different situations:

1. Very often, in particular with consumer data, the observed set is fixed and independent of $i$ and $t$: $J_{it} = J$. For instance, it can be $\{1,2,3\} = \{\text{beer, wine, cider}\}$ for every observations.

2. In controlled experiments, it can be designed differently. For instance, we can have observations for all pairwise choices, that is $J_{it}$ equal to $\{1,2\}$, $\{1,3\}$ and $\{2,3\}$. We can also have observations to find the effect of introducing a new alternative. For instance $\{1,2\}$, then $\{1,2,3\}$. 


Without observing all cases, i.e. \{1,2\}, \{1,3\}, \{2,3\} and \{1,2,3\}, the choice probabilities of the underlying RUM with context effects are not nonparametrically identifiable. In other words, without observing all cases, the extrapolation of the observed decisions to all possible choices is assumed in the model. For empirical applications, unless variations in the choice set are observed across time or locations, estimation of the model might prove to be difficult and possible only experimentally. In this paper, the empirical application to the deodorant market provided sufficient variations in choice sets to estimate our normalization parameters, but this might not always be the case. This limitation must be taken into consideration by managers seeking to optimize their product portfolio with this structural model.

4.4 Hierarchical Bayes Model

The model of Section 4.2 contains a large number of parameters even after taking into account the identifying restrictions. These unknown parameters will be analyzed in a Bayesian framework. We follow a hierarchical Bayesian approach with several layers:

- **Layer 1 (Section 4.2)** specifies the distribution of the observations \(y\) given \(X\) and parameter \(\theta_i\).
- **Layer 2 (Section 4.4.1)** specifies the distribution of \(\theta_i\) given hyperparameters. The distribution is assumed independent of \(X\).
- **Layer 3 (Section 4.4.2)** specifies the prior distributions on the set of hyperparameters.

4.4.1 Individual Heterogeneity

To facilitate computations, we decompose the weights \(w_{ik}\) into two sub-parameters \(v_{1ik} > 0\) and \(v_{2ik} > 0\) such that:

\[
w_{ik} = \frac{\nu_{1ik}}{\nu_{1ik} + \nu_{2ik}}\]

We assume parameters follow specific distributions in the population with hyperparameters to be estimated. For parameters restricted to be positive, we assume that they follow independent gamma distributions with shape \(\eta\) and rate \(\lambda\) (see Appendix B.1):

\[
r_{ik} \sim \gamma(\eta_k^r, \lambda_k^r)
\]

\[
\mu_{0ik} \sim \gamma(\eta_k^\mu, \lambda_k^\mu)
\]
\[\nu_{1ik} \sim \gamma(\eta_k^\nu_1, 1)\]
\[\nu_{2ik} \sim \gamma(\eta_k^\nu_2, 1)\]
\[c_i \sim \gamma(\eta_c^c, \lambda_c^c)\]
\[\sigma_{ik}^2 \sim \gamma(\eta_k^\sigma, \lambda_k^\sigma)\].

Then \(w_{ik}\) follows a Beta distribution (see Appendix B.3).

For the qualitative attributes, we assume that \(a_{ik} = (a_{ikl}, l = 1, ..., L_k)\) follows a Dirichlet distribution (see Appendix B.4):
\[a_{ik} \sim \text{Dirichlet}(d_{kl}, l = 1, ..., L_k)\].

The Dirichlet distribution normalizes the \(a_{ikl}\) such that \(\sum_{l=1}^{L_k} a_{ikl} = 1\). This normalization solves the identification problem described in Section 4.3.1.

The parameters \(\beta_{ik}\) can be either positive or negative. For these parameters, we define a multivariate normal distribution allowing for correlations:
\[(\beta_{i1}, ..., \beta_{iK})' \sim \text{MVN}(m, \Omega),\]
with mean \(m\), and variance \(\Omega\), a \((K \times K)\) positive definite matrix.

We regroup the hyperparameters under the notation \(\omega\) such that:
\[\omega = (\eta_k^\nu, ..., \eta_k^\nu, \lambda_{kl}^a, ..., \lambda_{kl}^a, d_{kl}, m, \Omega; k = 1, ..., K, l = 1, ..., L_k).\]

Now that we have defined the distribution to represent individual heterogeneity and we can marginalize out the individual level parameters. The likelihood becomes:
\[l(y|\omega, X) = \int l^*(y|\theta, X)g(\theta|\omega)d\theta,\]
where \(g(\theta|\omega)\) denotes the joint heterogeneity distribution.

### 4.4.2 Prior and Posterior Distributions

We use a hierarchical Bayes approach and define prior distributions on the hyperparameters. Hierarchical Bayes models are common in discrete choice models with heterogeneity (see Allenby and Rossi, 1998, 2006).

Since the priors can have a significant impact on the final result, especially when the number of observations is rather small compared to the number of hyperparameters, they have to be chosen following certain principles:
1. They have to be defined after having treated the identifiability issues discussed in Section 4.3.

2. They have to correspond to the expected range for the parameters. This point is especially important for the neural interpretation. For instance, if a variable is expected to be within a particular range, say $[0, 1]$, the parameters have to be normalized such that the variable falls approximately in that range. On the other hand, when parameters do not have particular restrictions to account for, a vague prior allowing for a wide range is chosen.

**Gamma($\eta, \lambda$) distributed parameters:** We use the following prior with density (see Appendix B.1.1):

$$p(\eta, \lambda) = \frac{1}{Z_0} \frac{p_\gamma^{\eta-1} e^{-\lambda q_\gamma} \gamma^{s_\gamma}}{\Gamma(\eta)^r_\gamma}, \eta > 0, \lambda > 0,$$

where $(p_\gamma, q_\gamma, r_\gamma, s_\gamma)$ are chosen by the researcher and $Z_0$ is a normalization constant ensuring that the density integrates to 1. The values for $(p_\gamma, q_\gamma, r_\gamma, s_\gamma)$ can be adjusted to make the hyperprior more or less vague depending on each case. We discuss briefly the values chosen for each application in Sections 5.2 and 6.2.

**Gamma($\eta, 1$) and corresponding Beta($\eta_1, \eta_2$) distributed parameters:** For the $\gamma(\eta', 1)$ distributed parameters $\nu_{1i}$ and $\nu_{2i}$, that compose the parameter $w_i$, we use a prior proportional to $\frac{0.4^{\eta-1}}{\Gamma(\eta)}$. The distribution for $\eta$ conditional on a gamma distributed individual level parameter $\nu_i$ has a density proportional to:

$$P(\eta|\nu_i) \propto \left(0.4 \prod_{i=1}^{N} \nu_i\right)^{\eta-1} \frac{1}{\Gamma(\eta)^{1+N}}.$$

The prior on $\eta$ was chosen such that the corresponding prior on $w = \frac{\nu_i}{\nu_1 + \nu_2}$ is roughly uniform on $[0, 1]$ with some additional mass very close to 0 and 1. A histogram of a Monte Carlo simulation of $w_i$ can be found in Figure 11.

**Dirichlet($d_{kl}, l = 1, \ldots, L_k$) distributed parameters:** For the Dirichlet distribution parameters, we use a prior where $d_{kl}$ are independent and $\gamma(1, 1)$ distributed:

$$P(d_k) = \prod_{l=1}^{L_k} P_\gamma(d_{kl})$$
$$= \prod_{l=1}^{L_k} e^{-d_{kl}}$$
$$= e^{-\sum_{l=1}^{L_k} d_{kl}},$$
where \( d_k = (d_{kl}, l = 1, ..., L_k) \). The distribution for \( d_k \) conditional on a Dirichlet distributed individual level parameter \( a_{ik} \) has a density proportional to:

\[
P(d_k|a_{ik}) \propto \left( \prod_{l=1}^{L_k} \Gamma(d_{l}) \right)^{N} e^{-\sum_{l=1}^{L_k} d_{kl}(1-\sum_{i=1}^{N} \log(a_{ik}))}.
\]

**Multivariate Normal \((m, \Omega)\) distributed parameters:** For the \( MVN(m, \Omega) \) distributed \( \beta_{ik} \) parameters, the conjugate distribution is the Normal-Inverse Wishart distribution:

\[
(m, \Omega) \sim NIW(0, 1, K + 4, 2I_K),
\]

where \( I_K \) is the identity matrix with \( K \) diagonal elements. The corresponding conditional distribution is:

\[
(n^\beta, \Omega|\beta) \sim NIW\left(\frac{N\bar{\beta}}{1+N}, 1+N, K+N+4, V_\beta\right)
\]

\[
V_\beta = 2I_K + \sum_{i=1}^{N} (\beta_i - \bar{\beta})(\beta_i - \bar{\beta})' + \frac{N}{1+N} \bar{\beta}\bar{\beta}'.
\]

**Posterior distributions:** Once the priors and the likelihood are specified, we can obtain the posterior distributions. The joint posterior distribution for the individual level parameters and hyper-parametrizes is given by:

\[
P(\omega, \theta|y, X) = \frac{l^*(y|\theta, X)g(\theta|\omega)\pi(\omega)}{\int \int l^*(y|\theta, X)g(\theta|\omega)\pi(\omega)d\theta d\omega},
\]

where \( \pi(\omega) \) denotes the prior density of \( \omega \). There exist two additional types of posterior distributions, corresponding to the parameters and the hyperparameters, respectively: \( P(\theta|y, X) \) and \( P(\omega|y, X) \).
4.4.3 Model Parameters and Hyperparameters

We present here a table with a summary of the model’s parameters and their distribution:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Domain</th>
<th>Number</th>
<th>Distribution</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>${\beta_{ik}}_{k=1}^{K}$</td>
<td>$R^K$</td>
<td>$N \times K$</td>
<td>$MVN(m, \Omega)$</td>
<td>Average preference weights</td>
</tr>
<tr>
<td>$r_{ik}$</td>
<td>$R_+$</td>
<td>$N \times K$</td>
<td>$\gamma(\eta^r_k, \lambda^r_k)$</td>
<td>Normalization scale</td>
</tr>
<tr>
<td>$\mu_{0ik}$</td>
<td>$R_+$</td>
<td>$N \times K$</td>
<td>$\gamma(\eta^\mu_k, \lambda^\mu_k)$</td>
<td>Prior beliefs</td>
</tr>
<tr>
<td>$w_{ik}$</td>
<td>$[0, 1]$</td>
<td>$N \times K$</td>
<td>$Beta(\eta_{ik}^{w1}, \eta_{ik}^{w2})$</td>
<td>Context weight</td>
</tr>
<tr>
<td>$c_i$</td>
<td>$R_+$</td>
<td>$N$</td>
<td>$\gamma(\eta^c, \lambda^c)$</td>
<td>Generalized mean parameter</td>
</tr>
<tr>
<td>$\sigma^2_{ik}$</td>
<td>$R_+$</td>
<td>$N \times K$</td>
<td>$\gamma(\eta^\sigma_k, \lambda^\sigma_k)$</td>
<td>Variance of taste for characteristic $k$</td>
</tr>
<tr>
<td>${a_{ikl}}_{l=1}^{L_k}$</td>
<td>$[0, 1]^{L_k}$</td>
<td>$N \times \sum_k L_k$</td>
<td>$Dirichlet({d_{kl}}_{l=1}^{L_k})$</td>
<td>Qualitative characteristics magnitude</td>
</tr>
</tbody>
</table>

4.5 Estimation Procedure

4.5.1 Simulation via Sequential Monte Carlo

We use the Sequential Monte Carlo (SMC) approach for hierarchical Bayesian estimation proposed by Daviet (2019) to simulate posterior distributions. We simulated the posterior distribution using 512 particles. The quality of the simulated approximation was evaluated by running the algorithm several times and adjusting the number of simulated cycles to obtain a stable simulated posterior across runs.

5 Experimental Application to Credit Card Preferences

In this section we analyze the data obtained from a discrete choice experiment, inspired by Fiebig et al. (2010) where respondents choose credit cards based on their characteristics. We first describe the data, and provide details about the prior distributions. We then discuss the posterior distributions.

5.1 Data Description

We asked 102 undergraduate students from the Rotman Business School at the University of Toronto to answer between 70 and 90 questions. For the first 10 questions, they were asked to choose one credit card from a set with 2 options. For the subsequent questions, the sets varied in size from 3 to 4 options. We have collected a total of 7,980 observations. When the subjects were asked 90
questions, the last 20 questions where randomly selected from the set of questions already asked to check for consistency. The individuals were informed that some questions might be repeated. Each credit card had 7 quantitative attributes described in the instruction form as follows\(^1\) (Fiebig et al., 2010, see):

- **Annual fee in Canadian dollars (between 5 and 25, with increments of 1):** The annual cost of the credit card’s ownership.

- **Transaction fee in Canadian dollars (between 0 and 0.5, with increments of 0.1):** Transaction (Payment / Withdrawal) charge.

- **Interest charged in % (between 5 and 25, with increments of 1):** The annual interest rate applied to the credit after the grace period of 30 days for each payment.

- **Reward in % (between 0 and 2.5, with increments of 0.25):** The % of each payment credited back at the end of the year.

- **ATM surcharge in Canadian dollars (between 0 and 2.5, with increments of 0.25):** The fee charged for ATM cash withdrawal.

- **Currency conversion fee in % (between 0 and 2.5, with increments of 0.25):** The additional % charged on a transaction in a foreign currency.

- **Warranty on purchase in days (0, 15, 30, 60, 90 or 120 days):** The card automatically insures new goods purchased with it for a certain number of days.

The respondents had an unlimited amount of time to answer each question.

\(^1\)The range of possible values is in parentheses and was not communicated before the experiment.
5.2 Hyperpriors

For the multivariate normal distributed parameters and the beta distributed parameters, we use the hyperpriors described in Section 4.4.2. Since all the attributes are quantitative, there are no Dirichlet distributed parameters.

For the gamma distributed parameters, we choose a diffuse hyperprior, with parameter values \((p_\gamma = 1, q_\gamma = 0, r_\gamma = 0, s_\gamma = 0)\). In that case, the prior is improper and uniform on \(\mathbb{R}^+\). The posterior distribution for \((\eta, \lambda)\) becomes proportional to the likelihood (see Appendix B.1.1).

5.3 Estimation Results

In this section we describe and discuss the simulated posterior distributions for each parameter and corresponding hyperparameters. Since the number of parameters is large relative to the number of observations per individual, several of the parameters may be estimated with relatively large variance. We expect this to occur especially for the hyperparameters. An approximation of the distribution moments is evaluated by computing the moments of the simulation draws.

We first provide an overview of the simulated aggregate distributions \(P(\theta | y, X)\) and \(P(\omega | y, X)\) for the parameter \(\theta\) and the corresponding hyperparameter \(\omega\). The former is the marginal posterior distribution of the individual-level parameters, and the latter is the posterior distribution for the hyperparameter. The simulated marginal posterior distribution \(P(\theta | y, X)\) is obtained by aggregating the simulated posterior distributions of the individual-level parameters for each individual, and treating it as one large sample.

The variance of the marginal posterior distribution \(P(\theta | y, X)\) can have two main sources. First, it can be the result of a large uncertainty on the parameters at the individual level. Second, it can originate from the variability across individuals. We propose a measure of the individual heterogeneity and discuss it in Section 5.3.2.

Finally we discuss in more details the posterior distributions for each parameter and comment on individual level distributions. We begin with the synaptic strength parameters \(\beta\) and \(\sigma\) as the identification of the other parameters depends on their sum being different from zero. We then move on to comment on the context weight parameters \(w\) and other valuation function parameters.
5.3.1 Population Level Posterior Distributions Overview

We report below the simulated moments for the hyperparameter distributions and the parameters’ marginal distributions, that is for the parameter $\theta$ and the corresponding hyper-parameter $\omega$, the simulated moments of $P(\omega|y,X)$ and $P(\theta|y,X)$. We present the results for relevant hyperparameters in Table 1. The summary statistics of the parameters’ marginal distributions are presented in Table 3.

We can note two important findings from these statistics. First, the mean values for the attribute preference weights $\beta$ deviated from their prior mean of 0 substantially for 4 out of 7 attributes. This indicates that there is evidence in the data for non-zero $\beta$ values and should allow for identification of the various valuation function parameters. Second, the posterior distribution of the $c$ parameter is very concentrated relative to the diffuse prior. Since this parameter was unidentified under the hypothesis of no context effect, there is strong evidence that the context matters in the attribute valuation process.
Table 1: Simulated posterior mean and standard deviation for hyperparameters. A "-" indicates the use of an improper diffuse prior.
\[ E(\Omega|y, X) \text{ and } SD(\Omega|y, X) \]

\[
E(\Omega_{kk}) = 0.67 \quad SD(\Omega_{kk}) = 0.94
\]

\[
E(\Omega_{kk'}) = 0.00 \quad SD(\Omega_{kk'}) = 0.58
\]

<table>
<thead>
<tr>
<th>YearFee</th>
<th>TransFee</th>
<th>Interest</th>
<th>Reward</th>
<th>ATMFee</th>
<th>CurrEx</th>
<th>Warranty</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.41</td>
<td>0.37</td>
<td>-1.51</td>
<td>0.41</td>
<td>-0.09</td>
<td>-0.10</td>
<td>0.13</td>
</tr>
<tr>
<td>(2.63)</td>
<td>(0.86)</td>
<td>(1.26)</td>
<td>(0.57)</td>
<td>(0.27)</td>
<td>(0.27)</td>
<td>(0.21)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TransFee</th>
<th>Interest</th>
<th>Reward</th>
<th>ATMFee</th>
<th>CurrEx</th>
<th>Warranty</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.81</td>
<td>-1.25</td>
<td>1.08</td>
<td>0.36</td>
<td>0.00</td>
<td>-0.08</td>
</tr>
<tr>
<td>(0.89)</td>
<td>(0.68)</td>
<td>(0.46)</td>
<td>(0.24)</td>
<td>(0.15)</td>
<td>(0.21)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interest</th>
<th>Reward</th>
<th>ATMFee</th>
<th>CurrEx</th>
<th>Warranty</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.17</td>
<td>1.68</td>
<td>0.34</td>
<td>0.08</td>
<td>-0.10</td>
</tr>
<tr>
<td>(2.08)</td>
<td>(0.35)</td>
<td>(0.08)</td>
<td>(0.05)</td>
<td>(0.07)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reward</th>
<th>ATMFee</th>
<th>CurrEx</th>
<th>Warranty</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.08</td>
<td>0.11</td>
<td>-0.03</td>
<td>-0.08</td>
</tr>
<tr>
<td>(0.68)</td>
<td>(0.46)</td>
<td>(0.23)</td>
<td>(0.21)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ATMFee</th>
<th>CurrEx</th>
<th>Warranty</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.36</td>
<td>-0.03</td>
<td>-0.08</td>
</tr>
<tr>
<td>(0.24)</td>
<td>(0.23)</td>
<td>(0.08)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CurrEx</th>
<th>Warranty</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.02</td>
<td>-0.10</td>
</tr>
<tr>
<td>(0.10)</td>
<td>(0.05)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Warranty</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
</tr>
<tr>
<td>(0.04)</td>
</tr>
</tbody>
</table>

Table 2: Simulated posterior mean and standard deviation for \( \Omega \).
Table 3: Simulated marginal posterior mean and standard deviation for individual level parameters.

<table>
<thead>
<tr>
<th>$\theta^*$</th>
<th>Attribute</th>
<th>Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>YearFee</td>
<td>TransFee</td>
</tr>
<tr>
<td>$\beta_k$</td>
<td>-3.06</td>
<td>-1.71</td>
</tr>
<tr>
<td></td>
<td>(3.45)</td>
<td>(1.96)</td>
</tr>
<tr>
<td>$\sigma_k^2$</td>
<td>0.37</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>(0.24)</td>
<td>(0.28)</td>
</tr>
<tr>
<td>$w_k$</td>
<td>0.48</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>(0.31)</td>
<td>(0.32)</td>
</tr>
<tr>
<td>$\mu_{0k}$</td>
<td>0.39</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>(0.28)</td>
<td>(0.25)</td>
</tr>
<tr>
<td>$r_k$</td>
<td>0.66</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>(0.54)</td>
<td>(0.72)</td>
</tr>
</tbody>
</table>

$\mu = 0.60 \ (0.26) \ -$
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>YearFee</td>
</tr>
<tr>
<td>$\beta_k$</td>
<td>0.81</td>
</tr>
<tr>
<td>$\sigma_k^2$</td>
<td>0.30</td>
</tr>
<tr>
<td>$w_k$</td>
<td>0.50</td>
</tr>
<tr>
<td>$\mu_{0k}$</td>
<td>0.35</td>
</tr>
<tr>
<td>$r_k$</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Table 4: Measure of individual heterogeneity for each parameter using the ratio of the standard deviation of the posterior means at the individual level to the overall posterior standard deviation.

We observe that a significant amount of $SD(\theta|y, X)$ is explained by individual heterogeneity for every parameter. In particular, most of the variance in $\beta$ is due to individual heterogeneity. Given the fact that i) $SD(\theta|y, X)$ is large, and ii) a significant share of it is due to individual heterogeneity, we conclude that there is substantial heterogeneity in the population for every parameter.

5.3.3 Posterior for $\beta$ and $\sigma$

We present a histogram for the simulated posterior distributions $P(m_k|y, X)$ and $P(\beta_k|y, X)$ for each attribute.

Figure 13: Simulated posterior distributions $P(m_k|y, X)$ and $P(\beta_k|y, X)$ for each attribute.

We present a histogram for the simulated posterior distributions $P(m_k|y, X)$ and $P(\beta_k|y, X)$ in Figure 13. The ratio of the posterior to the prior distributions can be found in the Appendix (Figure 32). The posterior means have expected signs, with negative means for cost variables (Yearly fees,
transaction fees, interest rates, ATM fees, and currency exchange fees) while desirable variables have positive means (Reward percentage and warranty). The variables ATM fees, currency exchange fees and warranty have means close to 0 indicating that they matter very little relative to other variables for most individuals.

| $\rho(\beta|y,X)$ | YearFee | TransFee | Interest | Reward | ATMFee | CurrEx | Warranty |
|-------------------|---------|----------|----------|--------|--------|--------|----------|
| YearFee           | 1.00    |          |          |        |        |        |          |
| TransFee          | 0.05    | 1.00     |          |        |        |        |          |
| Interest          | -0.17   | -0.23    | 1.00     |        |        |        |          |
| Reward            | 0.10    | -0.20    | 0.31     | 1.00   |        |        |          |
| ATMFee            | -0.05   | -0.13    | 0.21     | 0.16   | 1.00   |        |          |
| CurrEx            | -0.06   | -0.00    | -0.02    | -0.03  | 0.25   | 1.00   |          |
| Warranty          | 0.09    | -0.11    | -0.06    | 0.02   | -0.05  | -0.17  | 1.00     |

Table 5: Simulated posterior marginal correlation of $\beta$’s. The prior correlation across attributes is 0.

We now discuss the posterior correlations between tastes for attributes $\rho(\beta|y,X)$ presented in Table 5. We observe that most correlations in tastes are weak, with a few exceptions:

- Individuals who care more about transaction fees tend to care less about other factors such as interest ($\text{corr} = -0.23$) and reward rate ($\text{corr} = -0.20$).

- Individuals who care more about interest, tend to also care more about reward rate ($\text{corr} = 0.31$) and ATM fee ($\text{corr} = 0.21$).

- There is also correlation between sensitivity to ATM fees and currency exchange fees ($\text{corr} = 0.25$). However, the currency exchange coefficient is low for most individuals.

From the correlations we suspect that there might be two groups of individuals, those mostly sensitive to transaction fees, and those sensitive to a mix of other attributes. The yearly fee sensitivity seems to affect most individuals and hence no strong correlation is detected with other attributes.

Concerning the variability in tastes for characteristics, we present histograms for the simulated posterior distributions $P(\sigma_k|y,X)$ in Figure 14. Interestingly, the mean values for all attributes
Figure 14: Simulated posterior distributions $P(\sigma_k | y, X)$ for each attribute.

are of comparable magnitudes, ranging from 0.58 to 0.63. Moreover, the possibility that $\sigma_k = 0$ has been ruled out for every attribute. This indicates that it is important to model the stochastic nature of the tastes for characteristics (at the synaptic level). A second important point is that even for attributes that matter very little in terms of preference weights, the attribute values are very relevant in computing choice probabilities due to their role in creating correlations in utility shocks.

5.3.4 Posterior for $w$

Figure 15: Simulated posterior distributions $P(\eta_{k1} | y, X)$ and $P(\eta_{k2} | y, X)$ for each attribute.

We present the histograms of the simulated posterior distribution for the hyperparameters $P(\eta_{k1} | y, X)$ and $P(\eta_{k2} | y, X)$ in Figure 16. We can observe a large deviations from the prior distributions for the hyperparameters. The probability mass is concentrated slightly below 1 for most of the attributes. This suggests that $w_{ik}$ is nearly uniform in the population, with some additional mass near 0 and 1. We had reported that the heterogeneity was high, with $He(w_k)$ ranging from 0.50 to 0.56. This suggests a high variability of $w$ across individuals and attributes. We thus need
Figure 16: Simulated posterior distributions $P(w_k|y, X)$ for each attribute.

to examine individual-level distributions to gain further insights.

Figure 17: Simulated posterior distributions of $E_i(w_{ik}|y, X)$ for each attribute (top). The prior values are $E_i(w_{ik}) = 0.50$ and $SD(E_i(w_{ik})) = 0.20$. The ratio of the posterior to the prior distributions is provided at the bottom.

We are mostly interested in the individual-level posterior distributions $P_i(w_{ik}|y, X)$. We report the simulated distributions of $E_i(w_{ik}|y, X)$, that is the distribution of individual posterior means, in Figure 17. From these distributions, we can verify the presence of substantial individual heterogeneity, with individual posterior means ranging from 0 to 1. An individual with a mean $w_{ik}$ close to 0 shows almost no context effect, while an individual with a mean of 1 assigns strong importance to the context. We might wonder if the context weight varies across attributes for a given individual. To address this question, we computed the correlations between $E_i(w_{ik}|y, X)$ for various attributes and reported them in Table 6. We see that overall these correlations are weak, indicating a variable importance of context within individuals across attributes.
\[ \rho(E_i(w_{ik}|y, X)) \]

<table>
<thead>
<tr>
<th></th>
<th>YearFee</th>
<th>TransFee</th>
<th>Interest</th>
<th>Reward</th>
<th>ATMFee</th>
<th>CurrEx</th>
<th>Warranty</th>
</tr>
</thead>
<tbody>
<tr>
<td>YearFee</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TransFee</td>
<td>0.01</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interest</td>
<td>0.07</td>
<td>-0.12</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reward</td>
<td>-0.09</td>
<td>-0.11</td>
<td>0.07</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATMFee</td>
<td>-0.04</td>
<td>0.02</td>
<td>-0.02</td>
<td>0.09</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CurrEx</td>
<td>0.06</td>
<td>-0.09</td>
<td>0.04</td>
<td>0.03</td>
<td>-0.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Warranty</td>
<td>0.00</td>
<td>-0.22</td>
<td>0.07</td>
<td>0.09</td>
<td>0.02</td>
<td>-0.15</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 6: Simulated correlations of individual-level posterior mean \( E_i(w_{ik}|y, X) \), across attributes.

In summary, the results indicate that although in general the context matters, its importance varies across individuals and attributes. This might indicate that individuals have strong beliefs about the distribution of characteristics for some attributes and weak beliefs for others.

### 5.3.5 Posterior for \( \mu_0 \)

![Histograms of \( P(\mu_{0k}|y, X) \) for each attribute.](image)

Figure 18: Simulated posterior distributions \( P(\mu_{0k}|y, X) \) for each attribute.

We present a histogram for the simulated posterior distributions \( P(\mu_{0k}|y, X) \) in Figure 18. We note that the distribution has a lot of mass near very low values, with a mean around 0.4.

### 5.3.6 Posterior for \( r \)

We present a histogram for the simulated posterior distributions \( P(r_k|y, X) \) in Figure 19. The posterior distributions are similar across attributes with a mean ranging from 0.37 to 0.98, and a mode around 0.3. While we cannot rule out the possibility that \( r = 1 \), in which case the model
would simplify as this parameter becomes irrelevant, it seems that $r$ is between 0 and 1. Since $r$ is used as a power, this would indicate a concave function.

### 5.3.7 Posterior for $c$

We present a histogram for the simulated posterior distribution of $P(c|y, X)$ in Figure 20. The posterior distribution has a mean of 0.60 (against 1.4 in the prior). This indicates that the reference point might be constructed using an average between a geometric mean and an arithmetic mean.

### 5.3.8 Context Effects

Our model predicts context effects when $\beta_k \neq 0$, $w_k > 0$ and $c$ is not too large. The results indicate that there is a strong heterogeneity across subjects and attributes. In these conditions, we expect context effects to be observed, but only for some subjects and on specific attributes. These attributes sensitive to context vary among the subjects. This has important implications for marketing practitioners working with context effects in advertising and sales. The context effects
are likely to be effective only on some types of consumers. Suitable segmentation strategies taking into account the main types of consumers are one possible practical strategy.

6 Empirical Application to the Deodorant Market

6.1 Data description

We apply this model to a data set detailed in Bronnenberg, Kruger, and Mela, 2008, known as the IRI Academic Data Set. The data set contains an unbalanced panel data on households purchases of various products. We focus this analysis on deodorant purchases in small stores and grocery stores for the year of 2012. This constitutes a sample size of 1,337 individuals for a total of 3521 observations.

![Choice set size](image)

Figure 21: Empirical distribution of the choice set size.

One of the limitations of this dataset is that we do not have inventory data for the small stores. We infer a part of the inventory for these stores at a given point of time by looking at recorded purchases in the store starting from the Monday of the week before and until the Sunday of the week after. If a product in inventory has not been sold in this period of time, or if the purchases were not recorded, the product would be missing from our database. We argue that this problem should be minimal as we doubt a store would keep in inventory a product that does not sell well. Concerning grocery stores, we have the complete inventory of products sold in a given week, resulting in larger choice sets. Overall, our average choice set is of 44.30 products with a standard deviation of 37.66. This is considerably larger than the experimental choice set size. We present a histogram of the empirical distribution of the choice set size in Figure 21.

The deodorants possess a total of 11 attributes, including 3 quantitative attributes and 8 qual-
itative attributes. The attributes and their descriptions are provided in the Table 7.

<table>
<thead>
<tr>
<th>Attribute (code)</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price (PRC)</td>
<td>Quantitative</td>
<td>Price value in dollars</td>
</tr>
<tr>
<td>Display (DSP)</td>
<td>Qualitative</td>
<td>Indicates if the product display was enhanced: minor or major (lobby/end aisle).</td>
</tr>
<tr>
<td>Feature (FEA)</td>
<td>Qualitative</td>
<td>Indicates if the product was featured: frequent shopper program only (such as additional points), small ad, medium ad, large ad, retailer coupon or rebate.</td>
</tr>
<tr>
<td>Price reduction (RED)</td>
<td>Qualitative</td>
<td>Indicates if the product benefited a price reduction of greater than 5%.</td>
</tr>
<tr>
<td>Brand (BRD)</td>
<td>Qualitative</td>
<td>One category for each of the 14 most sold brand, and minor brands regrouped under the label &quot;other&quot;.</td>
</tr>
<tr>
<td>Anti-perspirant (PSP)</td>
<td>Qualitative</td>
<td>Indicates if the product has anti-perspirant properties</td>
</tr>
<tr>
<td>Form (FRM)</td>
<td>Qualitative</td>
<td>Solid, stick, spray, gel, clear, wise, invisible, roll, and other. A product can belong to several categories (e.g. &quot;clear solid stick&quot;).</td>
</tr>
<tr>
<td>Scent (SCT)</td>
<td>Qualitative</td>
<td>A product belongs to one or more categories presented in Table 8.</td>
</tr>
<tr>
<td>Strength (STR)</td>
<td>Quantitative</td>
<td>Effective time if indicated (24 or 36 hours), 0 otherwise.</td>
</tr>
<tr>
<td>Aluminum (ALM)</td>
<td>Qualitative</td>
<td>Indicates if the product contains aluminum salts.</td>
</tr>
<tr>
<td>Size (SIZ)</td>
<td>Quantitative</td>
<td>Product weight measured in ounces.</td>
</tr>
</tbody>
</table>

Table 7: Deodorants’ attributes description

6.2 Hyperpriors and Priors

For the beta distributed parameters, we use the hyper-priors described in Section 4.4.2.

For this application we split the $\beta_k$ into two groups, one with support on $\mathbb{R}$, and one restricted to the positive (or negative) domain. The coefficients $\beta$ for the preferences on antiperspirant, strength, aluminum and size follow a multivariate normal distribution. We used the prior described in Section 4.4.2. The coefficient for the price is such that its additive inverse follows a gamma distribution,

$\beta_k$.
as we wanted to restrict it to the negative domain: $-\beta_{PRC} \sim \gamma(\eta_{PRC}^\beta, \lambda_{PRC}^\beta)$. The coefficients for display, featured, and price reduction follow a gamma distribution to restrict them to the positive domain. Finally, to solve an identification issue that we discuss below, the coefficients for brand, form and scent are also restricted to be on the positive domain and follow a gamma distribution.

The identification problem arises for qualitative attributes where the number of characteristics $L_k > 1$ and where there exists no item possessing none of the characteristics. Indeed, suppose we have a qualitative attribute $k$ where the characteristics’ values are ordered as $a_1 < a_2 < \ldots < a_{L_k}$ and $\beta_k > 0$. The values ordering $a_1 > a_2 > \ldots > a_{L_k}$ with $\beta_k < 0$ yields the same preference ordering. For this reason, we use gamma distributed $\beta$ parameters for these qualitative attributes. These attributes are the brand, the form, and the scent.

<table>
<thead>
<tr>
<th>Scent family</th>
<th>examples</th>
<th>Scent family</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fresh</td>
<td>frost, polar</td>
<td>Texture</td>
<td>cotton, silk</td>
</tr>
<tr>
<td>Water/elements</td>
<td>ocean, storm</td>
<td>Romance</td>
<td>love, magnetic</td>
</tr>
<tr>
<td>Active/energy</td>
<td>dynamic, adrenaline</td>
<td>Effectiveness</td>
<td>dry, control</td>
</tr>
<tr>
<td>Flower</td>
<td>lily, blossom</td>
<td>Paradise</td>
<td>exotic, fiji</td>
</tr>
<tr>
<td>Sweet/fruity</td>
<td>vanilla, cherry</td>
<td>Emotions/abstract</td>
<td>ambition, brave</td>
</tr>
<tr>
<td>Plants</td>
<td>tea, forest</td>
<td>Standard</td>
<td>classic, regular</td>
</tr>
<tr>
<td>Clean</td>
<td>clean, pure</td>
<td>Other</td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Description of deodorant scent families

For every gamma distributed parameter, denoted $\theta^*$ here, we choose a shrinkage hyperprior,
with parameter values \((p_\gamma = 0.5, q_\gamma = 1.5, r_\gamma = 0.1, s_\gamma = 0.4)\). The prior is chosen to solve the dimensionality problem relative to the large number of parameters. The prior allows for potentially high values of the individual level parameter, while maintaining density around 0. Using this prior, if the posterior shifts the probability mass away from 0, it is the result of information from the data. We provide in Figure 22 a histogram representing the marginal distributions of \(\eta, \lambda\) and \(\theta_i^*\).

### 6.3 Estimation Results

The population-level statistics suggest a very strong heterogeneity across individuals. The population-level posterior distributions can be found in Appendix D.2. We focus our analysis on the individual-level posterior distributions and show that the estimation method is informative at the individual level. In the following sections, we will thus discuss the posterior distributions at the individual level for randomly selected individuals for each parameter. The conclusions provided generalize to samples of other individuals. This highlights the relevance of our method for individual preference estimation widely used in marketing and advertising.

#### 6.3.1 Posterior for \(\beta\) and \(\sigma\)

At the population level, the distributions are relatively spread out, indicating substantial heterogeneity in tastes. We present histograms for the population level posterior distributions in Appendix D.2. The correlations between \(\beta_k\) across attributes are negligible (largest correlation of 0.07).
Figure 23: Histograms representing the simulated posterior distributions of $P_i(\beta_{ik}|y,X)$ (MVN) for 3 randomly selected individuals (1 per row). The ratio of the posterior to the prior distribution is represented with a red dashed line.

Figure 24: Histograms representing the simulated posterior distributions of $P_i(\beta_{ik}|y,X)$ (Gamma) for 3 randomly selected individuals (1 per row). The ratio of the posterior to the prior distribution is represented with a red dashed line.
To visualize the heterogeneity we provide posterior distributions for a subsample of 3 randomly selected individuals (Figure 23). This allows to confirm substantial heterogeneity in tastes. We can however note that the individual-level posterior distributions are more informative than the population level posterior distributions. The few observations per individual precludes accurate estimation of the magnitude of $\beta_{ik}$. However, the sign is identified for several cases. This allows us to estimate the parameters from the value function. Heterogeneity can also be found in the gamma distributed $\beta_{ik}$ (Figure 24).

![Histograms representing the simulated posterior distributions of $P_i(\sigma_{ik}|y,X)$ for 3 randomly selected individuals (1 per row). The ratio of the posterior to the prior distribution is represented with a red dashed line.](image)

Concerning the variability in tastes for characteristics, the population level posterior means for each $\sigma_{ik}$ are considerably larger than in the experimental application, with values ranging from 1.57 to 1.63. Moreover, the possibility that $\sigma_k = 0$ is negligible for every attribute. A similar pattern can be found at the individual level (Figure 25). The heterogeneity seems less important than for other parameters. This confirms that it is important to model the stochastic nature of the tastes for characteristics (synaptic plasticity). We also confirm that even for attributes that matter very little in terms of preference weights, the attribute values are still relevant in computing choice probabilities due to their role in creating correlations in utility shocks.

As for the experimental application, the probability that $(\beta_k + \epsilon_k) = 0$ is very low. This should
allow for identification of the value function parameters.

6.3.2 Posterior for $w$

Figure 26: Simulated posterior distributions of $E_i(w_{ik}|y, X)$ for each attribute (top). The prior values are $E_i(w_{ik}) = 0.50$ and $SD(E_i(w_{ik})) = 0.20$. The ratio of the posterior to the prior distributions is provided at the bottom.

Figure 27: Histograms representing the simulated posterior distributions of $P_i(w_{ik}|y, X)$ for 3 randomly selected individuals (1 per row). The ratio of the posterior to the prior distributions is represented with a red dashed line.
We present in Figure 26 the distribution of the individual level posterior means \( E_i(w_{ik}|y, X) \) in the population. We can see that individuals have widely varying posterior means, and that the posterior has a higher standard deviation than the prior. A lot of mass from the prior distribution has been shifted toward 0 and 1 in the posterior distribution. This confirms presence of a high degree of heterogeneity across individuals.

Examining several individual-level posterior distributions (Figure 27), we note that heterogeneity is present not only across individuals, but also across attributes. Within an individual, the context might be extremely important for some attributes (\( w \) is close to 1), and of almost no importance for others.

This implies that tangible context effects might occur only for some attributes and that these attributes vary across individuals.

### 6.3.3 Posterior for \( \mu_0 \)

We present a histogram for the simulated population level posterior distributions \( P(\mu_{0k}|y, X) \) in the Appendix (Figure 39). We note that these distributions are relatively spread out with means ranging from 2.98 to 3.80.
Figure 28: Histograms representing the simulated posterior distributions of $P_i(\mu_{0ik}|y, X)$ for 3 randomly selected individuals (1 per row). The ratio of the posterior to the prior distributions is represented with a red dashed line.

Examining the individual level distributions, we can also detect heterogeneity. We have to consider that $\mu_{0ik}$ is only identified when $w_{ik} < 1$. Since $w_{ik}$ might be close to 1 in several cases, the posterior does not always strongly differ from the prior. In our subsample, the lowest posterior mean is 0.51 and the highest is 7.43.
6.3.4 Posterior for $r$

Figure 29: Histograms representing the simulated posterior distributions of $P_i(r_{ik}|y,X)$ for 3 randomly selected individuals (1 per row). The ratio of the posterior to the prior distributions is represented with a red dashed line.

The posterior distributions for $r_{ik}$ remain spread out over a large range of values. A dataset with more observations per individual might be necessary to estimate $r_{ik}$ with higher accuracy.
6.3.5 Posterior for $c$

Figure 30: Histograms representing the simulated posterior distributions of $P_i(c_i|y,X)$ for 9 randomly selected individuals (3 per row). The ratio of the posterior to the prior distributions is represented with a red dashed line.

We present a histogram for the simulated posterior distribution of $P(c|y,X)$ in the Appendix (Figure 40). The posterior mean is around 2.8, which is higher than the prior mean (2.45) and the posterior mean for $c$ in the deodorant experiment (0.60). This might indicate that in presence of large choice sets, individuals give more importance to the top values for each attributes.

When looking at the individual level posterior distributions (Figure 30), we notice that there might be two types of individuals. One type has a value for $c$ between 0 and 1, and might create a reference point based on a geometric or arithmetic mean. This is similar to the results from the experimental analysis. Another type of individual, has larger values for $c$, and might use a reference point that is based on the maximum value found in the choice set. In that case, the context effects might disappear as adding lower value alternative would not change the reference point.
6.3.6 Posterior for $a$

The population level posterior distributions remain very similar to the prior distributions (see Appendix, Figures 41 and 42). As for other parameters, this is mostly due to individual heterogeneity.

When looking at a subsample of individuals, we confirm that individuals have various preferences (see Figure for an example using the display attribute). A similar pattern can be found for every categorical attribute. A few of the categories have a posterior with mass on higher values while most of the other categories lose in value. This diversity of tastes justifies the large panel of choices provided by stores.

![Histograms representing simulated posterior distributions $P_i(a_{ik}|y, X)$ using the display attribute as an example for 3 randomly selected individuals (1 per row). The ratio of the posterior to the prior distributions is represented with a red dashed line.](image)

Figure 31: Histograms representing simulated posterior distributions $P_i(a_{ik}|y, X)$ using the display attribute as an example for 3 randomly selected individuals (1 per row). The ratio of the posterior to the prior distributions is represented with a red dashed line.

6.3.7 Context Effects

When analyzing simultaneously the posterior distributions for $\beta_{ik}$, $w_{ik}$ and $c_i$, we find results comparable to the experimental application of our model. In particular, we conclude that context matters for some individuals on some attributes, but that there is a strong heterogeneity across
both individuals and attributes.

In particular, it appears that some individuals have strong prior beliefs for given attributes, making the normalization process insensitive to the context. In addition, the potentially large values of $c$ for some individuals indicate that a part of the population generates reference points based on a maximum instead of an average. The resulting reference point is insensitive to the presence of lower valued attributes. For these individuals, the model predicts that no context effect will occur. This might be a consequence of the large choice set sizes encountered in practice.

We also identified several individuals for whom the reference point is generated based on a statistic close to a geometric or arithmetic mean. For these individuals, the model predicts that context effects can be strong. An appropriate segmentation could allow marketing practitioners to leverage context effects on these context sensitive segments.

7 Conclusion

The role of normalization, a canonical neural computation, has been recognized as essential in the process of decision and valuation. In this paper, we have introduced a new decision model for multi-attribute discrete choice capturing such normalization. We establish that when applied to a multi-attribute discrete choice problem, normalization explains the main context dependencies observed in the empirical literature, namely the asymmetric dominance effect, the compromise effect and the similarity effect. The empirical results support that context, through the normalization mechanism, plays an important role in the decision process. However, the statistical analysis reveals a high degree of heterogeneity. More precisely, the weight of the context in the attribute valuation process varies both across individuals and attributes. Moreover, empirical data reveals that the reference point is not always generated from an average, but that a maximum is used instead by some individuals. This explains the heterogeneity observed in the empirical literature on context effects. Future research will undoubtedly account for the heterogeneity and use appropriate methods such as hierarchical Bayes models with clustering.

In addition, we provide evidence that the tastes for various product characteristics vary over time and can be perceived as stochastic. This could be explained biologically through the role of synaptic plasticity. The main consequence is in the noise structure of the option valuations, where correlations between utilities depend on the attribute values. This supports similar findings concerning the correlations structures in the error term, and provides a biological explanation for
it (Dotson et al., 2017).

Overall, this paper contributes to the growing literature supporting the use of neuroscientific insights for development of economic theory. This approach allows for a better understanding of the processes involved in economic decisions and provides powerful modeling tools. Future research might focus on comparing SAN with other models allowing for the existence of context effects (Bhatia, 2013; Roe, Busemeyer, and Townsend, 2001; Trueblood, Brown, and Heathcote, 2014; Usher and McClelland, 2001).

References

Daviet, R. (2019). “Sequential Monte Carlo for Hierarchical Bayes with Large Datasets”. In: Available at SSRN 3382624.


Miller, R. B. (1980). “Bayesian Analysis of the Two-Parameter Gamma Distribution”. In: Techno-
metrics 22.1, pp. 65–69.
Noguchi, T. and N. Stewart (2014). “In the attraction, compromise, and similarity effects, alter-
natives are repeatedly compared in pairs on single dimensions”. In: Cognition 132.1, pp. 44–
56.
on human choice behaviour”. In: Nature Communications 7, pp. 1–11.
Schwartz, O. and E. P. Simoncelli (2001). “Natural signal statistics and sensory gain control”. In:
— (2014). “Vices and Virtues of Misguided Replications: The Case of Asymmetric Dominance”.
In: Journal of Marketing Research 51.4, pp. 514–519.
sion”. In: Journal of Marketing Research 29.3, pp. 281–295.
cumulator model of context effects in multialternative choice”. In: Psychological Review 121.2,
179–205.
pp. 281–299.
Divisive Normalization and Neurobiological Constraints on Choice”. In: Under Review.
A Appendix: Properties of the Subjective Value Function

A.1 Proof: \( \frac{\partial \Delta v(x_{1k}, x_{2k}, \mu_k)}{\partial \mu_k} \) has a root at \( \mu_k = \sqrt{x_{1k}x_{2k}} \)

We have shown that the derivative of \( \Delta v(x_{1k}, x_{2k}, \mu_k) \) is:

\[
\frac{\partial \Delta v(x_{1k}, x_{2k}, \mu_k)}{\partial \mu_k} = r \mu_k^{r-1} \left( \frac{x_{1k}^r}{(x_{1k}^r + \mu_k^r)^2} - \frac{x_{2k}^r}{(x_{2k}^r + \mu_k^r)^2} \right).
\]

Consequently, it admits a root at the point satisfying:

\[
\frac{x_{1k}^r}{(x_{1k}^r + \mu_k^r)^2} = \frac{x_{2k}^r}{(x_{2k}^r + \mu_k^r)^2}
\]

\( \Leftrightarrow \)

\[
x_{1k}^r (x_{2k}^r + \mu_k^r)^2 = x_{2k}^r (x_{1k}^r + \mu_k^r)
\]

\( \Leftrightarrow \)

\[
\mu_k^r (x_{2k}^r/2 - x_{1k}^r/2) = x_{1k}^r (x_{2k}^r/2 - x_{1k}^r/2)
\]

\( \Leftrightarrow \)

\[
\mu_k = \sqrt{x_{1k}x_{2k}}.
\]

A.2 Solutions of \( \Delta v(x_{1k}, x_{2k}, \mu_k) = b \)

We want to show that there exist only two values of \( \mu_k \) giving the same value for \( \Delta v(x_{1k}, x_{2k}, \mu_k) \) (except at the maximum) and that they satisfy the relation \( \mu_k^{(1)} = x_{1k} x_{2k}/\mu_k^{(2)} \).

First it is important to note that \( v(x_{jk}, \mu_k) \) satisfies the relation:

\[
v(x_{jk}, \mu_k) = \frac{x_{jk}^r}{\mu_k^r + x_{jk}^r} = 1 - \frac{\mu_k^r}{\mu_k^r + x_{jk}^r}
\]

Then it is straightforward to show that:

\[
\frac{x_{2k}^r}{\mu_k^r + x_{2k}^r} - \frac{x_{1k}^r}{\mu_k^r + x_{1k}^r} = \frac{\mu_k^r}{\mu_k^r + x_{1k}^r} - \frac{\mu_k^r}{\mu_k^r + x_{2k}^r} = \frac{x_{2k}^r}{(x_{1k} x_{2k}/\mu_k)^r + x_{2k}^r} - \frac{x_{1k}^r}{(x_{1k} x_{2k}/\mu_k)^r + x_{1k}^r}
\]
Let us now show that \( \mu_k^{(1)} \) and \( \mu_k^{(2)} \) are the only two solutions. We want to solve the equation for \( \mu_k \):

\[
\left( \mu_k^{r_k} + x_{2k} \right) \left( \mu_k^{r_k} + x_{1k} \right) = \mu_k^{r_k} \left( \mu_k^{r_k} + x_{2k} \right) + \mu_k^{r_k} \left( \mu_k^{r_k} + x_{1k} \right)
\]

This is a equation of degree 2 in \( \mu_k^{r_k} \), hence it has at most 2 solutions that are the cases derived above.

**A.3 Proof:** \( \arg \max_{x_{3k}} \Delta v(x_{1k}, x_{2k}, \mu_k(X)) \geq 0 \) if \( \frac{x_{2k}}{x_{1k}} \leq \left( \frac{3+\sqrt{5}}{2} \right)^{2/c} \)

Without loss of generality, we assume \( x_{1k} < x_{2k} \). If it exists, the optimal value \( x_{3k}^* \) is:

\[
\arg \max_{x_{3k}} \Delta v(x_{1k}, x_{2k}, \mu_k(X)) = \left(3\sqrt{x_{1k}x_{2k}^c} - x_{1k}^c - x_{2k}^c \right)^{1/c},
\]

where \( X = \{x_{1k}, x_{2k}, x_{3k}\} \).

Let us check that the right-hand side of this equation is well defined. We have:

\[
\left(3\sqrt{x_{1k}x_{2k}^c} - x_{1k}^c - x_{2k}^c \right)^{1/c} \geq 0
\]

\[
\iff 3 \left( \frac{x_{2k}}{x_{1k}} \right)^{c} \left( \frac{x_{2k}}{x_{1k}} \right)^{c} - 1 \geq 0
\]

\[
\iff 3 \sqrt{z} - z^c - 1 \geq 0, \text{ with } z = \frac{x_{2k}}{x_{1k}}
\]

The polynomial \( 3 \sqrt{z} - z^c - 1 \) admits 2 roots at \( \left( \frac{3+\sqrt{5}}{2} \right)^{2/c} \approx 6.85^{1/c} \) and at \( \left( \frac{3-\sqrt{5}}{2} \right)^{2/c} \approx 0.38^{1/c} \), and is positive in between. Since \( z > 1 \), we can ignore the second root.

**A.4 Proof:** \( \arg \max_{x_{3k}} \Delta v(x_{1k}, x_{2k}, \mu_k(X)) < \max(x_{1k}, x_{2k}) \)

Without loss of generality, we assume \( x_{1k} < x_{2k} \). We know that if \( \frac{x_{2k}}{x_{1k}} \leq \left( \frac{3+\sqrt{5}}{2} \right)^{2/c} \), then:

\[
\arg \max_{x_{3k}} \Delta v(x_{1k}, x_{2k}, \mu_k(X)) = \left(3\sqrt{x_{1k}x_{2k}^c} - x_{1k}^c - x_{2k}^c \right)^{1/c}.
\]

We want to show that:

57
\[(3\sqrt{x_{1k}x_{2k}} - x_{1k} - x_{2k})^{1/c} < x_{2k}\]
\[\Leftrightarrow \left(3\sqrt{x_{2k} - \left(\frac{x_{2k}}{x_{1k}}\right)^c - 1} \right)^{1/c} < \frac{x_{2k}}{x_{1k}}\]
\[\Leftrightarrow (\sqrt{z}^c - z^c - 1)^{1/c} < z, \text{ with } z = \frac{x_{2k}}{x_{1k}}\]
\[\Leftrightarrow \sqrt{z^c} - z^c - 1 < z^c\]
\[\Leftrightarrow \sqrt{z^c} - 2z^c - 1 < 0\]  \(\text{(7)}\)

We note that \(z = 1\) is a root of \(\sqrt{z^c} - 2z^c - 1\), and that since \(x_{1k} \leq x_{2k}, z \geq 1\). If we can show that \(\sqrt{z^c} - 2z^c - 1\) is decreasing on \(z > 1\), the inequality \((7)\) will be satisfied on that domain. We thus want to show that:

\[
\frac{\partial}{\partial z} (3\sqrt{z^c} - 2z^c - 1) < 0
\]
\[\Leftrightarrow \frac{3}{2}cz^{\frac{c}{2} - 1} - 2cz^{c-1} < 0\]
\[\Leftrightarrow \frac{3}{2}z^{\frac{c}{2} - 1} < 2z^{c-1}\]
\[\Leftrightarrow \frac{3}{4}z^{-\frac{c}{2}} < 1\]

Since \(z > 1\), this is always true.

**A.5 Condition:** \(\Delta v(x_{1k}, x_{2k}, \mu_k(X_0)) \leq \Delta v(x_{1k}, x_{2k}, \mu_k(X_2))\) for \(x_{1k} < x_{2k}, X_0 = \{x_{1k}, x_{2k}, 0\}, \text{ and } X_2 = \{x_{1k}, x_{2k}\}\)

For convenience, in the following demonstration we will use \(z = \frac{x_{2k}}{x_{1k}}\).

We want to know when this equation is satisfied:
The polynomial \( z^2c + 1 - 4z^c \) admits 2 roots at \( z = (2 - \sqrt{3})^{1/c} \) and \( z = (2 + \sqrt{3})^{1/c} \). Note that the first root is less than 1 and can be ignored. Since the polynomial is negative between its roots, the condition is satisfied for \( z < (2 + \sqrt{3})^{1/c} \).

A.6 Proof: \( \left( \frac{x_{1k}^c + x_{2k}^c}{2} \right)^{1/c} > \sqrt{x_{1k}x_{2k}} \)

We want to show that the reference point in the binary case is higher than the reference point maximizing the distance between the normalized values of \( x_{1k} \) and \( x_{1k} \).

\[
\left( \frac{x_{1k}^c + x_{2k}^c}{2} \right)^{1/c} > \sqrt{x_{1k}x_{2k}}
\]

\[
\Leftrightarrow \quad x_{1k}^c + x_{2k}^c - 2\sqrt{x_{1k}x_{2k}}c > 0
\]

\[
\Leftrightarrow \quad (\sqrt{x_{1k}^c} - \sqrt{x_{2k}^c})^2 > 0
\]

which is true.
A.7 Proof: \( \left( \frac{x_{ck} + 2x_{2ck}}{3} \right)^{1/c} \) > \( \left( \frac{x_{ck} + x_{2ck}}{2} \right)^{1/c} \), if \( \frac{x_{2ck}}{x_{1k}} > 1 \) (similarity effect)

\[
\frac{2}{3} x_{1k} + \frac{4}{3} x_{2ck} > x_{1k} + x_{2ck}
\]

\[
\frac{1}{3} (x_{2ck} - x_{1k}) > 0
\]

which is true if \( \frac{x_{2ck}}{x_{1k}} > 1 \). The converse is also true: \( \left( \frac{x_{ck} + 2x_{2ck}}{3} \right)^{1/c} < \left( \frac{x_{ck} + x_{2ck}}{2} \right)^{1/c} \), if \( \frac{x_{2ck}}{x_{1k}} < 1 \).

B Appendix: Review of Distributions

B.1 Gamma distribution

The gamma distribution is a continuous distribution for positive random variables. A gamma distribution with shape parameter \( \eta > 0 \) and rate parameter \( \lambda > 0 \) has the following probability density function:

\[
p_{\gamma}(x|\eta, \lambda) = \frac{\lambda^\eta}{\Gamma(\eta)} x^{\eta-1} e^{-\lambda x},
\]

where \( \Gamma(\cdot) \) is the gamma function defined by \( \Gamma(\eta) = \int_0^\infty e^{-x} x^{\eta-1} dx \).

The gamma distribution’s mean and variance are:

\[
E[x] = \frac{\eta}{\lambda}, \quad Var[x] = \frac{\eta}{\lambda^2}.
\]

B.1.1 Conjugate Prior

The gamma distribution admits as a conjugate prior with density defined on \( \mathbb{R}^2_+ \):

\[
p(\eta, \lambda) = Z_0^{-\eta} e^{-\lambda q} \chi^{(\eta s)} \Gamma(\eta)^r, \quad \eta > 0, \ \lambda > 0,
\]

where \( (p, q, r, s) \) are positive parameters and \( Z_0 \) is a normalizing constant ensuring the density function integrates to 1.

The likelihood for a sample \( x_i, i = 1, ..., n \) of independent \( \gamma(\eta, \lambda) \) distributed variables is:

\[
\prod_{i=1}^n p(x_i|\eta, \lambda) = \frac{\lambda^\eta x^{\eta+n}}{\Gamma(\eta)^n} \prod_{i=1}^n x_i^{\eta-1} e^{-\lambda \sum_{i=1}^n x_i}.
\]

Consequently, the posterior distribution obtained from the conjugate prior and the likelihood is (Miller, 1980):

\[
p(\eta, \lambda|x_1, ..., x_n) = \frac{Z_1^{-\eta+n} (p \prod_{i=1}^n x_i)^{\eta-1} e^{-\lambda(q+n) \sum_{i=1}^n x_i} \lambda^{(s+n)}}{\Gamma(\eta)^{r+n}},
\]

60
where $Z_1$ is a normalizing constant ensuring the density function integrates to 1.

The conjugate prior with parameters $(p = 1, q = 0, r = 0, s = 0)$ is an improper diffuse prior that is uniform on $\mathbb{R}_+^2$. In that case the posterior density is proportional to the likelihood.

### B.2 Multivariate Normal Distribution (MVN)

The MVN distribution is a continuous distribution defined over the space of a $K$-dimensional variable $x = (x_k; k = 1, ..., K)$, $x \in \mathbb{R}^K$. A MVN distribution with mean $m = (m_1, ..., m_K)$ and $K \times K$ positive definite Variance-Covariance matrix $\Omega$ has the following probability density function:

$$p_{\text{MVN}}(x|m, \Omega) = \frac{1}{2\pi |\Omega|^{0.5}} e^{-0.5(x-m)'\Omega^{-1}(x-m)},$$

where $|\cdot|$ denotes the determinant.

### B.3 Beta Distribution

The beta distribution, denoted $\text{Beta}(a, b)$, is a continuous distribution for random variables defined on the interval $[0, 1]$. A beta distribution with shape parameters $a > 0$ and $b > 0$ has the following probability density function:

$$p_{\beta}(x|a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1,$$

where $\Gamma(\cdot)$ is the gamma function.

#### B.3.1 Additional Results

If two independent variables $\nu_{1i}$ and $\nu_{2i}$ have respective distributions $\gamma(\eta_1, 1)$ and $\gamma(\eta_2, 1)$, then $w_i = \frac{\nu_{1i}}{\nu_{1i} + \nu_{2i}}$ follows a $\text{Beta}(\eta_1, \eta_2)$ distribution. The probability of observing $\nu_1 = (\nu_{1,1}, ..., \nu_{1,n})$ is as follows (or for $\nu_2 = (\nu_{2,1}, ..., \nu_{2,n})$, without loss of generality):

$$p(\nu_1|\eta_1) = \frac{1}{\Gamma(\eta_1)^n} \left( \prod_{i=1}^n \nu_{1i} \right)^{\eta_1-1} e^{-\sum_{i=1}^n \nu_{1i}}.$$

We can define a distribution on $\eta_1$ and $\eta_2$ such that:

$$p(\eta_1) = \frac{1}{Z_1} \frac{0.4^{\eta_1-1}}{\Gamma(\eta_1)} \quad \text{and} \quad p(\eta_2) = \frac{1}{Z_2} \frac{0.4^{\eta_2-1}}{\Gamma(\eta_2)}, \quad \eta_1 > 0, \eta_2 > 0,$$

where $Z_1$ and $Z_2$ are normalization constants ensuring the probabilities integrate to 1. In that condition, the posterior distribution of $\eta_1$ (and $\eta_2$) is as follows:

$$p(\eta_1|\nu_1) = \frac{1}{Z'_{1}} \frac{1}{\Gamma(\eta_1)^{n+1}} \left( 0.4 \prod_{i=1}^n \nu_{1i} \right)^{\eta_1-1} e^{-\sum_{i=1}^n \nu_{1i}}, \quad \eta_1 > 0,$$
where $Z'_1$ is a normalization constant ensuring the probability integrates to 1.

We can consider a population where each individual $w_i$ follows a $Beta(\eta_1, \eta_2)$ distribution. In that case, the expected $w_i$ for a given individual is:

$$E[w_i|\eta_1, \eta_2] = \frac{\eta_1}{\eta_1 + \eta_2}.$$ 

Assuming $\eta_1$ and $\eta_2$ are distributed as previously defined, we want to know the a priori distribution on individual means of $w_i$:

$$p(E[w_i|\eta_1, \eta_2]) = p\left(\frac{\eta_1}{\eta_1 + \eta_2}\right).$$

To derive this distribution, we proceed to a change of variable and define two variables $u$ and $v$ such that:

$$u = \frac{\eta_1}{\eta_1 + \eta_2}, \quad v = \eta_1 + \eta_2,$$

$$\eta_1 = u v, \quad \eta_2 = v (1 - u).$$

From this definition it follows that:

$$p(u, v) = \left| \begin{array}{cc} \frac{\partial \eta_1}{\partial u} & \frac{\partial \eta_1}{\partial v} \\ \frac{\partial \eta_2}{\partial u} & \frac{\partial \eta_2}{\partial v} \end{array} \right| p(\eta_1) p(\eta_2)$$

$$= \left| \begin{array}{cc} v & u \\ -v & (1 - u) \end{array} \right| p(\eta_1) p(\eta_2)$$

$$= (\eta_1 + \eta_2) p(\eta_1) p(\eta_2)$$

$$= (\eta_1 + \eta_2) \frac{1}{Z_1 Z_2} \frac{0.4^{\eta_1 + \eta_2 - 2}}{\Gamma(\eta_1) \Gamma(\eta_2)}$$

$$= v \frac{1}{Z_1 Z_2} \frac{0.4^{v-2}}{\Gamma(u v) \Gamma(v(1 - u))}.$$ 

We now just need to integrate $v$ out to obtain the desired distribution:

$$p\left(\frac{\eta_1}{\eta_1 + \eta_2}\right) = p(u)$$

$$= \int_0^\infty p(u, v) dv$$

$$= \frac{1}{Z_1 Z_2} \int_0^\infty v \frac{0.4^{v-2}}{\Gamma(u v) \Gamma(v(1 - u))} dv.$$ 

This integral can easily be numerically approximated.
B.4 Dirichlet Distribution

The Dirichlet distribution is defined over the space of an $L$-dimensional positive variable $a = (a_l; l = 1, ..., L)$ such that $\sum_{l=1}^{L} a_l = 1$. It admits $L$ positive parameters $d = (d_l; l = 1, ..., L)$ called concentration parameters. The probability density function is:

$$P_D(a|d) = \frac{\Gamma \left( \sum_{l=1}^{L} d_l \right)}{\prod_{l=1}^{L} \Gamma(d_l)} \prod_{l=1}^{L} a_l^{d_l-1}.$$  

The mean and variance of the $a_l$’s are respectively:

$$E[a_l] = \frac{d_l}{d_0}, \quad Var[a_l] = \frac{d_l(d_0 - d_l)}{d_0^2(d_0 + 1)},$$

where $d_0 = \sum_{l=1}^{L} d_l$.

When all the parameters are equal, say $d_l = d^*$, the distribution is symmetric. When $d^*$ is large, the joint probability density concentrates around $a_l = 1/L, l = 1, ..., L$. When $d^* = 1$, the density is uniform over the space of $a$.

The beta distribution is a special case of the Dirichlet distribution where $a$ has only 2 dimensions.

B.5 Inverse Wishart Distribution

The inverse Wishart distribution is a continuous distribution over positive-definite matrices $X$ of dimension $p$. It admits a positive-definite matrix $\Psi$ as the scale parameter and a positive scalar $\nu > p - 1$ as parameter for the numbers of degrees of freedom. Its probability density function is:

$$f(X; \Psi, \nu) = \frac{|\Psi|^{\nu/2}}{2^{\nu p/2} \Gamma_p(\nu/2)} |X|^{-(\nu+p+1)/2} e^{-\nu/2 \text{tr}(\Psi X^{-1})},$$

where $\Gamma_p(\cdot)$ is the multivariate gamma function.

The mean of the distribution is:

$$E[X] = \frac{\Psi}{\nu - p - 1}, \quad v > p + 1.$$  

The variance of each element of $X$ is:

$$Var[x_{ij}] = \frac{(\nu - p + 1)\Psi_{ij}^2 + (\nu - p - 1)\Psi_{ii}\Psi_{jj}}{(\nu - p)(\nu - p - 1)^2(\nu - p - 3)}, \quad v > p + 3.$$  

The covariance between elements of $X$ is:

$$Cov[x_{ij}, x_{kl}] = \frac{2\Psi_{ij}\Psi_{kl} + (\nu - p - 1)(\Psi_{ik}\Psi_{jl} + \Psi_{il}\Psi_{kj})}{(\nu - p)(\nu - p - 1)^2(\nu - p - 3)}, \quad v > p + 3.$$
C Metropolis-Hastings within Gibbs

In our Sequential Monte Carlo approach, we use a Metropolis-Hastings within Gibbs sampling method for the mutation step.

We want to simulate draws \((\theta^{(s)}, \omega^{(s)})\) from a target distribution \(P(\theta, \omega|y, X)\). One popular approach to achieve this goal in hierarchical Bayes models is to use Gibbs sampling. In this approach, one samples draws alternatively from the conditional distributions:

\[
\theta_i^{(s)} \sim P(\theta_i|\omega^{(s)}, y, X), \quad i = 1, \ldots, N,
\]

\[
\omega^{(s)} \sim P(\omega|\theta^{(s)}, y, X).
\]

Applying these steps a sufficient number of times, the distribution of \((\theta^{(s)}, \omega^{(s)})\) converges to the target distribution.

While sampling from \(P(\omega|\theta^{(s)}, y, X)\) is straightforward using conjugacy rules, the sampling step from \(P(\theta_i|\omega^{(s)}, y, X)\) requires additional consideration. For this purpose, we use a Metropolis-Hastings approach (Hastings, 1970; Metropolis et al., 1953). The use of Metropolis-Hastings steps within Gibbs steps has been discussed in details by Gilks, Best, and Tan, 1995.

The Metropolis-Hastings approach is a type of Markov Chain Monte-Carlo (MCMC) method to simulate draws from a distribution, in our case \(P(\theta_i|\omega^{(s)}, y, X)\). This method is based on an accept-reject approach where starting from an initial value \(\theta_i^{(s)}\), an alternative value \(\theta_i^{(s)'}\) is proposed. This alternative value replaces the original value with a probability computed as follows:

\[
P_{\text{accept}}(\theta_i^{(s)'}|\theta_i^{(s)}) = \frac{P(\theta_i^{(s)'}|\omega^{(s)}, y, X) \cdot q(\theta_i^{(s)} \rightarrow \theta_i^{(s)'})}{P(\theta_i^{(s)}|\omega^{(s)}, y, X) \cdot q(\theta_i^{(s)} \rightarrow \theta_i^{(s)'})}.
\]

where \(q(\theta_i^{(s)} \rightarrow \theta_i^{(s)'}\) is called the proposal distribution and determines how \(\theta_i^{(s)'}\) is chosen given \(\theta_i^{(s)}\). Applying Metropolis-Hastings accept-reject steps a sufficiently high number of times, the distribution of \(\theta_i^{(s)}\) will converge to \(P(\theta_i|\omega^{(s)}, y, X)\).

We are now going to describe the proposal distributions for each elements of \(\theta_i\). The proposal distributions are chosen to be independent such that:

\[
q(\theta_i^{(s)} \rightarrow \theta_i^{(s)'}) = q(\beta_i^{(s)} \rightarrow \beta_i^{(s)'}) \cdot q(\sigma_i^{2(s)} \rightarrow \sigma_i^{2(s)'})
\]

\[
\cdot q(w_i^{(s)} \rightarrow w_i^{(s)'}) \cdot q(\mu_0^{(s)} \rightarrow \mu_0^{(s)'})
\]

\[
\cdot q(r_i^{(s)} \rightarrow r_i^{2(s)'}) \cdot q(c_i^{(s)} \rightarrow c_i^{(s)'})
\]

\[
\cdot q(a_i^{(s)} \rightarrow a_i^{(s)'})
\]
For all gamma distributed variables $\theta^{(s)}_{\gamma}$, we use the following method to draw a proposal:

$$\theta^{(s)}_{\gamma}' = \theta^{(s)}_{\gamma} \cdot \epsilon_{\gamma},$$

where $\epsilon_{\gamma} \sim \gamma(100, 100)$. The proposed $\theta^{(s)}_{\gamma}'$ has thus an expected value of $\theta^{(s)}_{\gamma}$ and a standard deviation of $0.1 \cdot \theta^{(s)}_{\gamma}$. The probability of proposing the reverse move, that is proposing $\theta^{(s)}_{\gamma}$ starting from $\theta^{(s)}_{\gamma}'$, is the probability of drawing $\frac{1}{\epsilon_{\gamma}}$ from a $\gamma(100, 100)$ distribution. The ratios of proposal distributions is easily computed:

$$\frac{q(\theta^{(s)}_{\gamma} \rightarrow \theta^{(s)}_{\gamma}' \rightarrow \theta^{(s)}_{\gamma})}{q(\theta^{(s)}_{\gamma}' \rightarrow \theta^{(s)}_{\gamma} \rightarrow \theta^{(s)}_{\gamma}')} = \epsilon_{\gamma}^{-198} \cdot e^{100(\epsilon_{\gamma} - 1/\epsilon_{\gamma})}.$$

Note that we use the same approach for $w$ by drawing proposals for its components $\nu_1$ and $\nu_2$ which are gamma distributed variables.

For $\beta$, we use a multivariate normal distribution:

$$\beta^{(s)}' \sim MVN(\beta^{(s)}, 0.1I),$$

where $I$ is the identity matrix. Due to the symmetry of the multivariate normal distribution, the ratio of the proposal distributions is one:

$$\frac{q(\beta^{(s)} \rightarrow \beta^{(s)}')} {q(\beta^{(s)}' \rightarrow \beta^{(s)})} = 1.$$

Finally, for Dirichlet distributed variables $a_k = (a_{kl}, l = 1, ..., L_k)$, we use a Dirichlet proposal distribution such that:

$$a^{(s)}_k' \sim Dirichlet(100 \cdot a^{(s)}_k).$$

This allows to have most of the proposal density concentrated around $a^{(s)}_k$. However, there is no simplification to compute the ratio of the proposal densities and hence the pdf value for each has to be fully computed.
D Appendix: Additional Estimation Results

D.1 Experimental Application

Figure 32: Ratio of posterior to prior distribution for $m_k$ (top) and $\beta_k$ (bottom) for each attribute. The ratios are proportional to $P(y|m_k, X)$ and $P(y|\beta_k, X)$, respectively.

D.2 Empirical Application

Figure 33: Simulated posterior distributions $P(m_k|y, X)$ and $P(\beta_k|y, X)$ for coefficients following a multivariate normal distribution.
Figure 34: Ratio of posterior to prior distribution for $m_k$ (top) and $\beta_k$ (bottom) for each attribute. The ratios are proportional to $P(y|m_k, X)$ and $P(y|\beta_k, X)$, respectively.

Figure 35: Simulated posterior distributions $P(m_k|y, X)$ and $P(\beta_k|y, X)$ for coefficients following a gamma distribution. The second row represents the posterior to prior ratio.
Figure 36: Simulated posterior distributions $P(\sigma_k | y, X)$ for each attribute.

Figure 37: Simulated posterior distributions $P(\eta_{k1} | y, X)$ and $P(\eta_{k2} | y, X)$ for each attribute.

Figure 38: Simulated posterior distributions $P(w_k | y, X)$ for each attribute.
Figure 39: Simulated posterior distributions $P(\mu_0|y, X)$ for each attribute (top). The ratio of the posterior to the prior distributions is provided at the bottom.

Figure 40: Simulated posterior distribution $P(c|y, X)$. 
Figure 41: Simulated posterior distribution $P(a|y, X)$ for Dirichlet distributed parameters (Part 1). The ratio of the posterior to the prior distribution is represented by a red dashed line.
Figure 42: Simulated posterior distribution $P(a|y, X)$ for Dirichlet distributed parameters (Part 2). The ratio of the posterior to the prior distribution is represented by a red dashed line.