ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
Master Thesis Econometrics and Management Science:
Business Analytics and Quantitative Marketing

Estimating Cannibalization

Using Bayes Estimates from a Multinomial Probit Model and the GHK simulator

Author:
S.A. van Weede
471633

External Supervisors:
MSc. G.I.S. van Grinsven
MSc. S.A. Gieske

Supervisor:
Prof. dr. P.H. Franses

Second Assessor:
Prof. dr. D. Fok

March 31, 2020

Abstract

Cannibalization of own assortment is a general concern to many retailers when introducing new products. The focus of this study is estimating the cannibalization effect in terms of change in discrete choice probability pre- and post-introduction. We use an MCMC Gibbs sampler to estimate individual-specific brand preference and sensitivity to marketing decision variables specified by the Multinomial Probit (MNP) model. Estimates for these parameters are then used to simulate choice probabilities by means of the GHK simulator. Given that the MNP model notoriously suffers from identification issues, we compare two different identifying model specifications, each placing a different restriction on the covariance matrix of errors. The element-restricted model fixes just one parameter of the diagonal of this matrix, while the trace-restricted model fixes the trace of the matrix. We evaluate both methods in a simulation study and conclude that trace-restricted model outperforms the element-restricted model in overall model-fit, although the element-restricted model has slightly better predictive accuracy. We then apply the trace-restricted model to empirical sales data containing a new product introduction in the laundry detergent category. This approach allows one to observe changes in competitive structure as well as changes in sensitivities to marketing decision variables in the face of new product introductions.

Keywords: cannibalization, discrete choice modeling, Multinomial Probit, MCMC Gibbs sampler, GHK simulator, identification, trace-restriction, unobserved heterogeneity

Contents

1 Introduction

Product assortment strategy is a central concern to many retailers of fast-moving consumer goods. Retailers, both online and traditional, cater to a host of customers, each with different demands and price sensitivities. While some may only be interested in low-priced standard goods, others are willing to pay a premium for a more high-quality brand, a different flavor, or a special edition. In an effort to serve as many customers as possible, retailers may choose to expand their product assortment by introducing new brands and new product variants. However, it is widely accepted that product proliferation may come at a cost operationally (Broniarczyk & Hoyer, 2006). Having more product variety complicates inventory control and quality assurances and increases forecasting errors and overhead costs (Kim & Chhajed, 2000). As such, new products must be introduced with careful consideration.

In evaluating the success of such introductions, one must consider how much new demand they generate, but also to what extent this demand is poached from the existing, or incumbent products. This is also known as the cannibalization effect (Mason & Milne, 1994). Ignoring this effect may result in an overestimation of the new product's performance. The importance of considering the cannibalization effect has not been understated in the literature, but the effect itself has almost exclusively been viewed as undesirable. However, from a retailer's perspective, cannibalization may be part of a profit-maximization strategy when a new product introduction gains territory from an incumbent product with lower margins. Furthermore, it provides an opportunity to gain insight into customer preferences through their brand-switching behavior.

In essence, estimating cannibalization comes down to estimating consumer demand pre- and post-introduction. Demand, or marketing models, measure the change in demand as a function of marketing decision variables employed by a firm, or its competitors, to facilitate managers in their marketing strategies (Leeflang, Wittink, Wedel, & Naert, 2013). Among such models, a distinction is made between aggregate demand models, describing behavior at the market, store or brand level, and individual demand models. The latter has the advantage of gaining insight into the effect of marketing instruments on a consumer or household level. Furthermore, as consumer tastes continue to diversify, aggregated insights may no longer accurately describe consumer decision making (Allenby & Rossi, 1998). Especially in the face of new product introduction, dissagregate demand models may prove useful, as individual switching behavior can be exploited to gain further understanding of customers' latent preferences and sensitivities to marketing instruments (Leeflang et al., 2013).

At the disaggregate level, the behavior of a consumer or household is best described by discrete variables. The purchasing of a product at a certain price and discount among all other alternatives, is a discrete choice. Formally, discrete choice analysis is defined as "the modeling of choice from a set of mutually exclusive and collectively exhaustive alternatives", where the choice is expressed in choice probabilities (Ben-Akiva, Lerman, & Lerman, 1985). Many approaches to estimating such models exist, and can roughly be divided into two categories: classical methods and Bayesian methods. Here, we take preference to Bayesian methods, because classical approaches to inferring individual-specific parameters are more computationally demanding and yield only approximations (Allenby & Rossi, 1998).

In this paper, we make use of the Multinomial Probit Model (MNP), a discrete choice model, estimated by Bayesian analysis. The aforementioned advantage of disaggregate demand modeling is simultaneously one of its main complexities. Because we observe the effect of the marketing instruments on an individual level, one must account for unobserved heterogeneity across households (Fok & Franses, 2001). The MNP model estimated by Bayesian methods is especially well-equipped to account for this. Unfortunately, the MNP model suffers from identification issues, making some model parameters inestimable in the absence of identifying restrictions. In this paper we compare two specifications of the MNP model, both placing a different identifying restriction on the covariance matrix of errors. The first places a restriction on a single element of this matrix, while the second places a restriction on the trace of the matrix as first introduced by Burgette and Nordheim (2012). While comparing the two model-specifications, our goal is two-fold: 1) estimate brand preference and sensitivity to marketing instruments at the individual level in the face of a new product introduction and 2) estimate the choice probabilities pre- and post-introduction to gain insight into the extent of cannibalization. To our knowledge, this application of the MNP model has not yet been researched in previous literature.

The remainder of this paper is structured as follows. Section 2 provides a review of related work on cannibalization and discrete choice models. Section 3 discusses our methodology to obtaining parameter estimates for the marketing instrument sensitivities and the choice probabilities. In Section 4, we evaluate our methods in a simulation study. In Section 5, we apply our methods to empirical data containing sales data for laundry detergent provided by online retailer Picnic. Conclusions are presented in Section 6 and lastly, we discuss the limitations of this research and offer ideas for future research in Section 7.

2 Literature review

This chapter is split into two sections. In Section 2.1, we discuss several approaches to estimating the cannibalization effect, as proposed in the literature so far. In Section 2.2, we discuss related work on discrete choice modeling, our choice approach to estimating cannibalization.

2.1 Cannibalization

Current literature regarding the cannibalization effect can broadly be divided into two approaches. While one approach focuses on identifying situations in which cannibalization may occur, the other is concerned with quantifying the extent of cannibalization. The former is more normative in nature and provides ways to respond to situations for which the potential for cannibalization is high. For instance, Srinivasan, Ramakrishnan, and Grasman (2005) show that the likelihood of a new product cannibalizing the sales of incumbent product increases with the similarity of the features between these products. Similar findings are reported by Buday (1989) and Copulsky (1976). Alternatively, Mazumdar, Sivakumar, and Wilemon (1996) and Moorthy and Png (1992) present normative frameworks for the timing of new product introductions when the potential for cannibalization is high.

The other stream of research takes a more quantitative approach to estimating cannibalization effects. For instance, Mason and Milne (1994) estimate the cannibalization effect of new cigarette variants by identifying core and fringe consumers of each brand, based on a clustering algorithm. Pairwise cannibalization effects between different brands of cigarettes are calculated based on the extent to which niches overlap. This model however, is dependent on consumer survey data which may be costly to obtain.

Alternatively, making use of observed sales data, Fok and Franses (2004) propose a market-share attraction model to estimate the effect of a new brand introduction on competitive structure for laundry detergent brands. Their methodology enables to statistically test for various changes, including changes in market share and changes in marketing-mix sensitivities, pre- and post-introduction. Gielens (2012) builds on this approach in estimating to what extent private label introductions change national brands' market position. Here, the market share attraction model is extended by modelling each product introduction by a step dummy variable allowing for a level shift in attraction each time a product is introduced. As such, the attraction of a brand is a function of all product introductions, as well as the usual controls and marketing-mix variables.

Bijmolt, Van Heerde, and Pieters (2005) take matters a step further. In addition to estimating the effect of new product introductions by means of a multiplicative sales model, they propose an approach to optimizing a retailers' assortment. The authors predict sales instead of market share and use these estimates in a Gibbs sampler to tackle the optimization problem. The multiplicative sales model however, can also be used in a non-predictive setting. Van Heerde, Srinivasan, and Dekimpe (2010) specify a Vector Error Correction (VEC) model and use Bayesian techniques to estimate how much of the demand generated by new products stems from the incumbent brands. The VEC specification enables distinguishing short-term fluctuations from long-term effects.

Though all aforementioned approaches provide important insights into the effect of new product introductions on incumbent products, all are based on aggregate data. Aggregate data can be very useful in the absence of dissaggregate data for gaining an understanding of a market as a whole (Sheffi, Hall, & Daganzo, 1982). However, dissaggregate data, such as household panel data, facilitates the analysis of individual consumer choice behavior and enable the researcher to gain insight into consumers' latent preferences. Such data is best utilized by means of discrete choice models. Instead of modeling market share as a function of a brand's attraction, one models individual consumers' choice probabilities which may be aggregated in a later stage. Such an approach allows one to account for individuals' varying tastes, i.e. consumer heterogeneity.

2.2 Discrete Choice Models

Discrete choice models have been used for quite some time across many disciplines to model stated preference, as well as revealed preference (Keane, 1997). Amongst this class of models, the most commonly used is the Multinomial Logit (MNL) as introduced by McFadden et al. (1973). The MNL model accredits its wide usage to the relative ease with which it is estimated. However, the model makes some very strict assumptions regarding the patterns of substitution across alternatives (Train, 2009). Namely, the MNL model exhibits *independence from irrelevant alternatives* (IIA), meaning that the relative odds of choosing one alternative over the other is the same regardless of the other alternatives available. However, this assumption may not be appropriate in the case of consumer brand-choice behavior because some alternatives within a choice set are more similar than others (Chintagunta, 2001).

Several solutions have been proposed to alleviate the IIA assumption. For instance, the nested logit model does so, by imposing a hierarchical partitioning of the choice set based on predetermined assumptions (Ansari, Bawa, & Ghosh, 1995). The division of the choice sets in nested

partitions is what induces a correlation structure between alternatives, as products within groups are more strongly correlated than products across groups. However, dividing products into predetermined groups is not realistic, as the division is not always clear and thus may not properly capture the substitution effects (Nevo, 2000).

A model that offers a more flexible specification is the Multinomial Probit model (MNP) (Hausman & Wise, 1978). Though less restrictive, the model only started to gain traction when advances in simulation techniques and computing power increased. This is because the MNP model does not yield closed-form expressions for the integrals expressing choice probabilities, and therefore cannot be computed numerically. Several classical approaches avoiding the evaluation of the likelihood function have been proposed i.e. Method of Simulated Moments (McFadden, 1989) and Method of Simulated Scores (Hajivassiliou & McFadden, 1998). However, they require extremely large sample sizes to yield accurate estimates, given that the approaches rely on asymptotic approximations. Albert and Chib (1993) and McCulloch and Rossi (1994) were first to propose Bayesian sampling methods. They proposed using a Gibbs sampler to sample latent parameters alongside the model parameters, also known as *data augmentation*. Since then, Bayesian analysis of the MNP model has been the topic of much work; see e.g. Nobile (1998), Chib, Greenberg, Chen, et al. (1998), McCulloch and Rossi (2000), Imai and Van Dyk (2004).

More recently, Burgette and Nordheim (2012) have made contributions to this body of literature by proposing an innovative identification approach. Identification, as will be discussed in Section 3.3, is necessary as the latent parameters of the MNP model are scale- and level-invariant, and multiplying all latent parameters by a positive constant does not change the ordering. Identification often comes in the form of a restriction on the covariance matrix. Mostly, such a restriction is imposed on a single element of the matrix, but Burgette and Nordheim (2012) propose a restriction on the trace of the matrix instead.

In this paper, we compare the identification strategy of Burgette and Nordheim (2012) to that of Imai and Van Dyk (2004) (an element-restricted model), as we apply both strategies to an MNP panel model. Furthermore, our model specification allows for heterogeneity among individuals by using individual-specific regression parameters β_i. As neither the trace-restricted model, nor the element-restricted model by the aforementioned authors, allow for this, we consider that to be our theoretical contribution to the literature. The estimation of cannibalization serves as an empirical application of our method.

3 Methods

We commence by introducing the Multinomial Probit (MNP) model specification and our mode of parameter estimation in Section 3.1 and Section 3.2. In Section 3.3, we dive into the identification restrictions. Next, we elaborate on our method of obtaining the choice probabilities in Section 3.4, and lastly we propose methods of evaluation to assess the accuracy of our models in Section 3.5. As discussed previously, our model is specified at the individual level, and as such we may use individual or household interchangeably. The same applies to the use of product, article or brand. Implementation of all methods is done in Python.

3.1 Model Specification

We define the utility that customer i gains from product j at time t as follows:

$$U_{ijt} = X'_{ijt}\beta_i + \epsilon_{ijt}. \tag{1}$$

Here X_{ijt} represents product-specific covariates such as price and price promotions and also includes an intercept. The coefficient β_i represents the customer's sensitivity to the marketing decision variables as well as the customer's intrinsic brand preference. Ultimately, the consumer chooses the brand for which the utility is maximized:

$$y_{it} = \underset{j=1,..,J}{\operatorname{argmax}}\, U_{ijt} \tag{2}$$

The error terms ϵ_{ijt} follow a multivariate normal distribution such that $\epsilon_{it} = (\epsilon_{i1t}, ..., \epsilon_{iJt}) \sim N(0,\Sigma)$, where Σ is a J x J matrix and J is the number of alternatives. This specification leads to the MNP model. Given that choice modelling depends merely on the ordering of the utilities and not on their absolute value, it is convention to model the utility relative to a base alternative. Hence, we define the *relative* utility as follows:

$$\begin{aligned} U_{ijt} - U_{iJt} &= (X_{ijt} - X_{iJt})'\beta_i + (\epsilon_{ijt} - \epsilon_{iJt}) \\ \tilde{U}_{ijt} &= \tilde{X_{ijt}}'\beta_i + \tilde{\epsilon_{ijt}} \text{ for } j = 1, ..., J-1. \end{aligned} \tag{3}$$

Here $\tilde{X}_{ijt} = [I_{J-1}, D_i]$, where I_j is the identity matrix and D_i is a $(J-1)$ x L matrix of differences in marketing-mix variables between each alternative and the base alternative. The relative utility $\tilde{U}_{ijt}$ is the utility of brand i relative to the utility of brand J. The relative utility of brand J is thus equal to zero. Due to this transformation, the dimension of $\tilde{\epsilon_{it}} = (\tilde{\epsilon_{i1t}}, ..., \tilde{\epsilon_{iJt}})$ is now

$J-1$ instead of J. The distribution of $\tilde{\epsilon_{it}}$ remains multivariate normal, but the covariance matrix of error differences $\tilde{\Sigma}$ needs to be derived from the covariance matrix of errors Σ. Fortunately, this derivation is straightforward as $\tilde{\Sigma} = M_i'\Sigma M_i$, where M_i is a $J-1$ identity matrix with an additional columns of -1's as the ith column.

As stated in Equation 1, households may be heterogeneous in their intrinsic brand preference and sensitivity to marketing decision variables. We formulate β_i as random draws from the multivariate normal distribution:

$$\beta_i = \beta + \nu_i, \tag{4}$$

where $\nu_i \sim N(0,\Sigma_\beta)$. Σ_β is the L x L diagonal covariance matrix of β_i which allows the unobserved shocks that influence marketing mix sensitivities to be correlated. Here L includes the $J-1$ intercepts and the covariates.

3.2 Estimation

Given the utilities $\tilde{U}_{ijt}$, ..., $\tilde{U}_{iJt}$, the probability that household i purchases article j on occasion t is then given by:

$$Pr[\tilde{U}_{ijt} > \tilde{U}_{ikt} \;\; \forall k \neq j] \tag{5}$$

Given that the utility $\tilde{U}_{ijt}$ is normally distributed, the joint distribution of $\tilde{U}_{ijt}$ is a multivariate normal distribution with dimension T_i x $(J-1)$. Notice the subscript on the T_i, as the number of trips may vary between customers. To account for the household heterogeneity, we must integrate over β_i, resulting in an (T_i x $(J-1)$)-dimensional integral that needs to computed N times to obtain the following expression for the likelihood:

$$\ell(\text{data}|\theta) = \prod_{i=1}^{N} \int Pr[\tilde{U}_{ijt} > \tilde{U}_{ikt} \;\; \forall k \neq j] f_N(\beta_i|\beta, \Sigma_\beta) d\beta_i \tag{6}$$

It is clear that this expression is analytically intractable. We must therefore rely on sampling methods, and more specifically Gibbs sampling as described by Geman and Geman (1984). To briefly describe the idea behind Gibbs sampling, we consider two random variables: θ_1 and θ_2. Gibbs sampling proceeds by iteratively sampling θ_1^{m+1} from the full conditional posterior distribution $f(\theta_1^m|\theta_2^m)$ and using that value θ_1^{m+1} to sample θ_2^{m+1} from $f(\theta_2^m|\theta_1^{m+1})$. Given that we use the previous value to randomly generate the next sample value, we obtain a Markov Chain. After the Markov Chain converges, the simulated values $\theta_1^{(\text{m})}$ and $\theta_2^{(\text{m})}$ can be used as a sample from the joint posterior distribution $p(\theta_1, \theta_2|y)$. This distribution can then be used to obtain

posterior results, such as the posterior mean where $E_{\theta|y}[\theta] \approx \frac{1}{M}\sum_{m=1}^{M}\theta^{(m)}$. One drawback of this Markov Chain Monte Carlo (MCMC) sampling algorithm is that the draws obtained by the sampler by definition suffer from autocorrelation. We may circumvent this issue by the use of *thinning* i.e. saving only every tenth draw.

Gibbs sampling is especially convenient in the context of the MNP model because we can make use of the fact that conditional on the utilities $\tilde{U}_{ijt}$, ..., $\tilde{U}_{iJt}$, the MNP model is a standard Bayesian linear regression. As such, the latent utilities can be sampled alongside the model parameters. This is also known as data augmentation. The layers for the Gibbs sampler are stated in Algorithm 1.

Algorithm 1 Gibbs Sampler Multinomial Probit

1: Set starting values for β, $\tilde{\Sigma}$, Σ_β and $\tilde{U}_{ijt}$, ..., $\tilde{U}_{iJt}$, m = 0

2: Draw $\tilde{U}_{ijt} \mid Y_{it}, \beta_i, \tilde{\Sigma}, \tilde{U}_{ikt} \;\; \forall \;\; k \neq j$ independently for all $t = 1, ..., T_i$ and $i = 1, ..., N$

3: Draw Σ:

- Set $\tilde{e_{it}} = (\tilde{U}_{it} - X'_{it}\beta_i)$
- Draw Σ from $IW(\nu, \Phi)$ where $\nu = \lambda + \sum_{i=1}^{N} T_i$ and $\Phi = \Psi + \sum_{i=1}^{N}\sum_{t=1}^{T_i} \tilde{e_{it}}\tilde{e_{it}}'$

4: Draw β_i conditional on β, $\tilde{\Sigma}$, Σ_β and $\tilde{U}_{ijt}$, ..., $\tilde{U}_{iJt} \quad \forall i$

5: Draw β conditional on Σ_β, $\tilde{\Sigma}$ and β_i

6: Draw Σ_β conditional on β, $\tilde{\Sigma}$ and β_i

7: Set $m = m + 1$ and return to step 2 till convergence

Step 2 through 6 of this algorithm require full conditional posterior distributions in order to sample the model parameters. These conditional distributions are derived from the posterior distributions which are proportional to the parameter's prior distribution multiplied by the likelihood. We take flat priors for all parameters, with the exception of the prior distribution for β_i (given by Equation 4) and for $\tilde{\Sigma}$, for which we choose a proper, but weakly informative prior:

$$p(\tilde{\Sigma}) \sim IW(\lambda, \Psi) \tag{7}$$

Here, the scale matrix Ψ determines the position of the distribution in the parameter space, while the degrees of freedom λ signifies the certainty of the prior information in the scale matrix (Schuurman, Grasman, & Hamaker, 2016). To specify the least informative prior possible, one may set Ψ equal to c times the identity matrix, where c is some constant, and λ such that it is only slightly larger than the number of random parameters. This because the inequality

$\lambda \geq (p-1)$ must hold in order to obtain a proper posterior density. However, some specifications may yield unstable results if the degrees of freedom is too small. Alternatively, the larger the degrees of freedom, the stronger the influence of the prior. The choice of λ is thus a balancing act between prior influence and numerical stability. The prior specified in Equation 7 has two very attractive properties. First, it serves as a way of placing identifying restrictions on $\tilde{\Sigma}$, which will be further discussed in Section 3.3. Second, it ensures that the posterior distributions for the remaining parameters are proper as well (Paap & Franses, 2000). A brief summary of the derivation of all posterior distributions can be found in Appendix A.

From most of these posterior distributions we can directly sample draws, as they are quite standard i.e. the (Multivariate) Normal distribution and the Inverted Wishart distribution. However, sampling the latent utilities requires a bit more effort. For instance, the utility $\tilde{U}_{ijt}$ given all other utilities $\tilde{U}_{it,(-j)}$ follows a truncated normal distribution, where the truncation region is determined as follows:

$$\tilde{U}_{ijt} \begin{cases} > \max\left(\tilde{U}_{it,(-j)}, 0\right) & \text{if } y_i = j \\ < \max\left(\tilde{U}_{it,(-j)}, 0\right) & \text{if } y_i \neq j \end{cases} \tag{8}$$

The truncation ensures that the relative ordering of the utilities remains intact. Many modern software packages allow for direct sampling from the truncated normal distribution, relying on the inverse cumulative distribution function (CDF) technique (Devroye, 1986). Sampling x from $TN(a, b)$ is then achieved by calculating the inverse CDF of u, such that $x = \Phi^{-1}(u)$, where $u \sim U[\Phi(a), \Phi(b)]$. This approach requires the evaluation of three integrals, and can be slow to be computed as u tends to zero or one (Geweke et al., 1991). Instead, we employ a mix of accept-reject sampling and exponential sampling, as done by Jiao and van Dyk (2015). In short, if $\tilde{U}_{ijt}$ is truncated such that $\tilde{U}_{ijt} \geq u$ and $u \leq 0$, we employ accept-reject sampling from an unconstrained normal distribution until $\tilde{U}_{ijt} \geq u$. Alternatively, if $\tilde{U}_{ijt} \geq u$ and $u \geq 0$, we use the exponential rejection sampling method as proposed by Robert (1995). The same scheme is used when $\tilde{U}_{ijt} \leq u$, but with a slight modification, as $-\tilde{U}_{ijt} \leq -u$. The mean and variance for the truncated normal is specified as follows:

$$\begin{aligned} \mu_{ijt} &= x'_{ij}\beta_i + F'(\tilde{U}_{it,(-j)} - X_{i(-j)}\beta_i) \\ \tau^2_{ijt} &= 1/\sigma_{jj} \end{aligned} \tag{9}$$

Here $F = -\sigma_{jj}\gamma_{j,-j}$, where σ_{jj} denotes the (j, j) element of Σ^{-1} and $\gamma_{j,-j}$ is the j^{th} row of Σ^{-1}

with the j^{th} element removed.

3.3 Identification

Though Algorithm 1 is a clever way of estimating the parameter estimates without having to evaluate a high-dimensional integral, the algorithm is complicated by the matter of identification. As is discussed extensively in the literature, the MNP model suffers from identification issues due to the fact that utilities are scale- and level-invariant. Multiplying all utilities by a positive constant does not change the ordering of an individual's choice probabilities. The same goes for adding a constant to all utilities. Given that we only observe the index of the maximum of the utilities (not the utilities itself), and the ordering of the utilities is preserved by such transformations, some of the model parameters are unidentified and cannot be estimated. The MNP model therefore requires some form of restriction to ensure all parameters are identified.

Common practice is to set $\tilde{\Sigma}_{1,1}$ equal to unity. This identifying restriction reduces the number of parameters to be estimated from $J(J+1)/2$ to $J(J-1)/2$ while maintaining all economically relevant information. The removed parameters merely contain the scale and level of utility and are thus not relevant for choice behavior and do not need to be considered (Train, 2009). However, the choice of which category corresponds to the element of $\tilde{\Sigma}$ that is set to unity, may have a large effect on the posterior predicted choice probabilities. Instead, Burgette and Nordheim (2012) propose a *trace-restricted* covariance matrix Σ. By fixing the trace rather than one element of the covariance matrix, one is no longer forced to choose which alternative has unit variance. Furthermore, these authors have found that the MNP model with trace-restricted covariance provides stronger identification, more interpretable results and yields less volatile posterior predictions, in comparison to the element-restricted covariance matrix as proposed by Imai and Van Dyk (2004). Burgette and Nordheim (2012) only implement the trace restriction for models that assume the mean of β to be zero, while Imai and Van Dyk (2004) also estimate models where the mean of β is expected to be non-zero. However neither groups of authors allow for heterogeneity in the regression parameters. This paper contributes to this research by implementing the trace-restriction to an MNP model that allows for heterogeneity among customers as well as assuming $\mathrm{E}[\beta_i|\beta, \Sigma_\beta] \neq 0$. We compare this model, hereby named the trace-restricted model, to the model where only the first element of Σ is set to unity, hereby named the element-restricted model.

3.3.1 Working Parameters

Instrumental to both the approach of Burgette and Nordheim (2012) and Imai and Van Dyk (2004) is the concept of *working parameters*. These parameters are not identified given the data Y but are identified as the parameter space is expanded by means of data augmentation to (Y, U). The key use of the working parameter is to improve the rate of convergence in the data augmentation algorithm as presented in Algorithm 1. To illustrate how the working parameter can achieve this, consider the likelihood of the model parameters $\theta = (\beta_i, \Sigma)$:

$$\mathcal{L}(\theta|Y) \propto P(Y|\theta) = \int P(Y, U|\theta)\, dU \tag{10}$$

Given that the working parameter denoted by α is not identified given the data, the likelihood of θ can also be defined as:

$$\mathcal{L}(\theta|Y) = \mathcal{L}(\alpha, \theta|Y) \propto \int P(Y, U|\alpha, \theta)\, dU \,\forall\, \alpha \tag{11}$$

This property of α can be used in one of two ways. First, one may condition on any value of α, which in the context of the MNP model often occurs in the form of a constraint (i.e. $\tilde{\Sigma}_{1,1} = 1$). This is also known as *conditional* data augmentation. Instead of conditioning on α, one may opt to average over the prior distribution of α. Using Fubini's theorem, allowing to switch the order of integration, the likelihood can then be written as:

$$\mathcal{L}(\theta|Y) \propto \int \left[\int \underbrace{P(Y, U|\alpha, \theta) P(\alpha|\theta)\, d\alpha}_{= P(Y,U|\theta)} \right] dU \tag{12}$$

Note that the prior distribution of α is conditional on θ. Averaging on α is also called *marginal data augmentation*. Though both approaches may be used, marginal data augmentation has a slight computational advantage because $\int P(Y, W|\alpha, \theta) P(\alpha|\theta)\, d\alpha$ is expected to introduce more variance into the conditional distributions than $P(Y, W|\alpha, \theta)$, because α is integrated out. More expected variance, allows the sampler to take larger jumps which leads to faster convergence. Additionally, it is expected to lead to Markov chains with less autocorrelation (Meng & Van Dyk, 1999). In short, working parameters may increase the rate of convergence of the Markov chains, especially when used in a marginal data augmentation algorithm.

3.3.2 Marginal Data Augmentation

As previously mentioned, we use a trace-restricted covariance matrix as defined by Burgette and Nordheim (2012) and apply it to the marginal data augmentation method as defined by Imai

and Van Dyk (2004), while assuming individual specific regression parameters whose population mean is non-zero. As such, we modify Algorithm 1 to incorporate the working parameter α resulting in Algorithm 2. Note that the starred parameters are only intermediate values. They are re-scaled by the working parameters in step 4.

Algorithm 2 Marginal Data Augmentation

1: Set starting values for β, $\tilde{\Sigma}$, Σ_β and $\tilde{U}_{ijt}$, ..., $\tilde{U}_{iJt}$, m = 0

2: Draw $\tilde{U}^*_{ijt} \mid Y_{it}, \beta_i, \tilde{\Sigma}, \tilde{U}_{ikt} \ \forall \ k \neq j$ independently for all $t = 1, ..., T_i$ and $i = 1, ..., N$

3: Draw Σ:

- Draw α^2 from $p(\alpha^2|\beta, \Sigma, Y) = p(\alpha^2|\Sigma)$
- Set $\tilde{e_{it}} = \alpha(\tilde{U}^*_{it} - X'_{it}\beta_i)$
- Draw $\tilde{\Sigma}^*$ from $IW(\nu, \Phi)$ where $\nu = \lambda + \sum_{i=1}^{N} T_i$ and $\Phi = \Psi + \sum_{i=1}^{N} \sum_{t=1}^{T_i} \tilde{e_{it}} \tilde{e_{it}}'$

4: Reset parameters:

- Set $\alpha^2 = \frac{1}{p-1} \operatorname{tr}(\Sigma^*)$
- $\tilde{\Sigma} = \alpha^{-2}\Sigma^*$
- $U = \alpha^{-1}\tilde{e_{it}} + X'_{it}\beta_i$

5: Draw β_i conditional on β, $\tilde{\Sigma}$, Σ_β and $\tilde{U}_{ijt}$, ..., $\tilde{U}_{iJt}$ $\quad \forall i$

6: Draw β conditional on Σ_β, $\tilde{\Sigma}$ and β_i

7: Draw Σ_β conditional on β, $\tilde{\Sigma}$ and β_i

8: Set $m = m + 1$ and return to step 2 till convergence

The trace-restricted marginal data augmentation algorithm as described in Algorithm 2 differs only from that of Imai and Van Dyk (2004) in the second part of step 4. Instead of setting α^2 equal to the average of the trace of $\tilde{\Sigma}^*$, they set α^2 equal to the first element of $\tilde{\Sigma}$. Apart from this step, the methods are identical. Note that in Step 5 through 7, we do not marginalize out α. The original IVD algorithm does marginalize out α when updating β, but that step is only possible when we assume $\mathrm{E}[\beta_i|\beta, \Sigma_\beta] \neq 0$. However, we do not make that assumption here and thus we only partially apply marginalization (Step 2 through 4). Imai and Van Dyk (2004) argue that the performance of this algorithm is less than when taking a full marginalization approach, but given that we do not want to make the zero-mean assumption, we accept a slight reduction in performance.

Step 3 of the marginal data augmentation algorithm requires drawing α^2 from $p(\alpha^2|\Sigma)$. Given the prior specification of $\tilde{\Sigma}$ in Equation 7, the prior of α^2 is proportional to:

$$p(\alpha^2|\Sigma) \propto |\tilde{\Sigma}|^{(-\lambda+p+1)/2} \exp -\frac{1}{\alpha^2} \operatorname{tr}(\Psi\tilde{\Sigma}^{-1}) * (\alpha^2)^{-\lambda p/2+1} \tag{13}$$

whereby integrating out α yields the prior distribution for $\tilde{\Sigma^*}$. A benefit of this prior specification, as pointed out by Burgette and Nordheim (2012), is that the Metropolis-Hasting step, introduced in the original MNP sampler by Albert and Chib (1993), is made redundant. We can simply sample $\tilde{\Sigma^*}$ from a standard inverse Wishart distribution and obtain $\tilde{\Sigma}$ by dividing $\tilde{\Sigma^*}$ by the average of its trace.

Running Algorithm 2 M times after convergence, and averaging over the M draws then yields the posterior parameter estimates for $\theta = (\beta_i, \tilde{\Sigma}, \beta, \Sigma_\beta)$. Convergence of the Markov Chain is assessed by means of subjective inspection as well as the Geweke test statistic (Geweke et al., 1991). This test statistic assesses convergence by comparing the mean of the estimates from the first 10% of the chain excluding burn-in, to the mean of the estimates of the last 50% percent of the chain, correcting for the correlation between the draws. The assumption is that if the means are drawn from a stationary distribution, they must be equal. The Geweke statistics follows a standard normal distribution and for any value $< |2|$, we fail to reject the hypothesis that the means are equal and that the chain has converged.

3.4 Choice probabilities

After obtaining the posterior estimates $\theta = (\beta_i, \Sigma, \beta, \Sigma_\beta)$, we turn to calculating the choice probabilities as given by Equation 5. The integrals solving for these probabilities do not have a closed-form expression and we must therefore rely on simulation to approximate the probabilities instead. For this, we use one of the most commonly used simulators, the GHK simulator after Geweke (1989), Hajivassiliou and McFadden (1998) and Keane (1994). Throughout the literature, the GHK simulator has confirmed to be the most accurate among other simulators. In the following section, we briefly discuss the GHK simulation algorithm as described by Train (2009).

3.4.1 GHK Simulator

Crucial to the implementation of the GHK simulator is that the simulation of the probability that individual i purchases product j at time t, P_{ijt}, occurs on utility differences. When simulating P_{ijt}, one subtracts the utility of U_{ijt} from all other utilities, but when simulating P_{ikt},

one subtracts the utility of U_{ikt} from all other utilities. The assumption is that P_{ijt} is equal to the probability that all differenced utilities are negative. To illustrate how the GHK simulator operates, consider the example of simulating P_{ikt}. Simulating P_{ikt} requires taking utility differences relative to the k^{th} alternative and transforming the J x J matrix Σ to obtain the $(J-1)$ x $(J-1)$ matrix $\tilde{\Sigma} = M_k \Sigma M_k'$. Here M_k is a $(J-1)$ identity matrix with an additional column of -1's as the k^{th} column. This yields:

$$\begin{aligned}
&\tilde{U}_{ijt} - \tilde{U}_{ikt} = (\tilde{V}_{ijt} - \tilde{V}_{ikt}) + (\tilde{\epsilon}_{ijt} - \tilde{\epsilon}_{ikt}) \ \ \forall \ \ j \neq k \\
&\tilde{U}_{ijkt} = \tilde{V}_{ijkt} + \tilde{\epsilon}_{ijkt} \\
&\tilde{\epsilon}_{it} = \{\tilde{\epsilon}_{i1t}, ..., \tilde{\epsilon}_{iJt}\} \text{ for all alternatives but the } k^{\text{th}} \text{ alternative} \\
&\tilde{\epsilon_{it}} \sim N(0, \tilde{\Sigma})
\end{aligned} \tag{14}$$

Given the lower triangular matrix L_k, where $L_k L_k' = \tilde{\Sigma}$, we can write the model as follows:

$$L_k = \begin{bmatrix} c_{11} & 0 & \dots & \dots & 0 \\ c_{21} & c_{22} & 0 & \dots & 0 \\ c_{31} & c_{32} & c_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \end{bmatrix} \tag{16}$$

$$\begin{aligned}
&\tilde{U}_{it1k} = \tilde{V}_{it1k} + c_{11}\eta_1 \\
&\tilde{U}_{it2k} = \tilde{V}_{it2k} + c_{21}\eta_1 + c_{22}\eta_2 \\
&\tilde{U}_{it3k} = \tilde{V}_{it3k} + c_{31}\eta_1 + c_{32}\eta_2 + c_{33}\eta_3 \\
&\vdots
\end{aligned} \tag{15}$$

where $\eta_i' = \{\eta_{1i}, ..., \eta_{J-1i}\}$ are $(J-1)$ independent draws from the standard normal distribution. If alternative k is chosen, this alternative has the highest utility and differencing the utilities for the remaining alternatives against the k^{th} utility thus yields negative utilities. As such, the choice probability P_{ikt} can be calculated as follows:

$$\begin{aligned}
P_{ikt} &= \Pr\left(\tilde{U}_{itjk} < 0 \ \ \forall j \neq k\right) \\
&= \Pr\left(\eta_1 < \frac{-\tilde{V}_{it1k}}{c_{11}}\right) \\
&\times \Pr\left(\eta_2 < \frac{-\tilde{V}_{it2k} + c_{21}\eta_1}{c_{22}} \ \middle| \ \eta_1 < \frac{-\tilde{V}_{it1k}}{c_{11}}\right) \\
&\times \Pr\left(\eta_3 < \frac{-\tilde{V}_{it2k} + c_{31}\eta_1 + +c_{32}\eta_2}{c_{33}} \ \middle| \ \eta_1 < \frac{-\tilde{V}_{it1k}}{c_{11}}, \ \ \eta_2 < \frac{-\tilde{V}_{it2k} + c_{21}\eta_1}{c_{22}}\right) \\
&\times \dots
\end{aligned} \tag{17}$$

The second line of Equation 17 is straight-forward to compute as $\left(\eta_1 < \frac{-\tilde{V}_{it1k}}{c_{11}}\right) = \Phi\left(\frac{-\tilde{V}_{it1k}}{c_{11}}\right)$, where Φ is the standard normal cumulative distribution function. The subsequent probabilities are slightly more complex because we condition on previous values of η. Consider the probability in the third line of Equation 17, which is expanded on below:

$$\begin{aligned}
&\Pr\left(\eta_2 < \frac{-\tilde{V}_{it2k} + c_{21}\eta_1}{c_{22}} \middle|\ \eta_1 < \frac{-\tilde{V}_{it1k}}{c_{11}}\right) \\
&= \Pr\left(\eta_2 < \frac{-\tilde{V}_{it2k} + c_{21}\eta_1}{c_{22}} \middle|\ \eta_1 = \eta_1^r\right) \\
&= \Phi\left(\frac{-\tilde{V}_{it2k} + c_{21}\eta_1^r}{c_{22}}\right)
\end{aligned} \tag{18}$$

Here, we condition on the value η_1 drawn in the previous step to calculate probability that $\tilde{U}_{it2k} < 0$. This value is drawn from a truncated standard normal, bounded from above by $\frac{-\tilde{V}_{it1k}}{c_{11}}$, and is denoted by η_1^r. Next, one draws the value η_2^r and calculates the subsequent probabilities in a similar way for each part of Equation 17. This procedure is repeated R times for each choice probability P_{ijt}, after which we average over all draws such that the simulated probability $\hat{P}_{ijt} = \frac{1}{R}\sum_r \hat{P}_{ijt}^{(r)}$.

Given that the Gibbs sampler produces estimates for the covariance matrix of error differences $\tilde{\Sigma}$, and the GHK simulator operates on the covariance matrix of errors Σ, we have to transform $\tilde{\Sigma}$ back into Σ. This is done by taking the Cholesky decomposition of $\tilde{\Sigma}$ and adding a row and column of zeros as the J^{th} column and J^{th} row. Multiplying the resulting matrix with its transpose then yields Σ.

3.5 Model Evaluation

Once the choice probabilities have been estimated, we compare the accuracy of the element-restricted model and the trace-restricted model amongst each other, but also against some benchmark measures. For these benchmark measures, we replace the GHK simulated choice probabilities $\hat{P}_{ijt}^{u_{ijt}}$ by the purchase frequency of a product over the estimation period. First, we compare the log likelihood of each model, where the MNP log-likelihood is given by:

$$\ell(\theta|y,x) = \sum_{i=1}^{N} \log\left(\prod_{t=1}^{T_i}\prod_{j=1}^{J} \hat{P}_{ijt}^{u_{ijt}}\right) \tag{19}$$

where $u_{ijt} = I(y_{ijt} = \max_k y_{ikt})$. Here the likelihood function is evaluated in the posterior means.

A more Bayesian approach to evaluating model fit is the widely applicable information crite-

rion, or WAIC, as introduced by Watanabe (2010). Model selection in Bayesian analysis is mostly achieved by using Bayes factors and odds ratios. However, numerical computation of Bayes factors in Hierarchical models may be complicated. Instead, one may approximate Bayes factors by means of information criteria, such as the commonly used Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC). An advantage of the WAIC is that the posterior density may stray from the multivariate normal, which is often required by other information criteria approximating the Bayes Factor. Furthermore, the WAIC is especially useful for models in which the number of parameters is unclear, as is the case for Hierarchical models. For this research, we use the WAIC as defined by Gelman, Hwang, and Vehtari (2014):

$$\begin{aligned} \text{WAIC} &= -2\left(\sum_{i=1}^{n} \log\left(\frac{1}{M}\sum_{m=1}^{M} p(y_i|\theta^m)\right) - p_{\text{WAIC}}\right) \\ p_{\text{WAIC}} &= \sum_{i=1}^{n} V_{m=1}^{M}(\log p(y_i|\theta^m)) \end{aligned} \tag{20}$$

The second term, the posterior variance of the log predictive density calculated for each datapoint y_i and summed over all individuals, is an approximation of the number of unconstrained parameters in the model. Parameters that are estimated without any constraints or prior information yield a count of one, while fully constrained parameters yield a count of zero. If both the data and the prior contribute to the estimation of the parameter, the parameter count is an intermediate value (Gelman et al., 2014). In this manner, the WAIC includes a penalty term for the number of parameters to penalize overfitting.

Next, to evaluate the models' predictive accuracy we look at a hold-out set. This set is created by excluding each customers' final s trips from the original dataset. These trips are not used in parameter estimation. We then assess the predictive accuracy for each customer by evaluating the mean absolute error at a given time t:

$$AE_{it} = \frac{1}{j}\sum_{j=1}^{j} |u_{ijt} - \hat{P}_{ijt}| \tag{21}$$

The AE_{it} is averaged over the s trips for each customer to obtain the mean absolute error per customer MAE_i. The overall MAE is obtained by averaging MAE_i over all customers. To further assess predictive accuracy, we look at the percentage of correct hits, and the log predictive likelihood. Here, the percentage of correct hits is defined as the percentage of times the sampler correctly assigns the highest choice probability to the chosen product.

4 Simulation study

During the simulation study, we evaluate the performance of both model specifications on simulated data. This data is simulated to resemble the data from our empirical application as much as possible. We discuss our simulation procedure in Section 4.1, and present our results directly after in Section 4.2.

4.1 Data and Prior Distributions

For a sample of $N = 100$ simulated customers, we simulate between 10 and 15 trips per customer. The number of choice alternatives is set to $J = 4$, each having $L = 2$ covariates excluding intercepts. These covariates are drawn with replacement from the set of price and price promotion data from the empirical dataset used in the following section. The intercepts for each individual $\boldsymbol{\beta}_i$ are generated from:

$$\begin{aligned} &\beta_{i1} \sim U(-0.75, -0.25), && \beta_{i2} \sim U(-0.3, 0.3) \\ &\beta_{i3} \sim U(-0.1, 0.4), && \beta_{i4} \sim U(0.0, 0.5) \\ &\beta_{i5} \sim U(-1.5, -0.5), && \beta_{i6} \sim U(0.25, 0.5) \end{aligned} \tag{22}$$

where β_{i1} through β_{i4} are the intercepts, and β_{i5} and β_{i6} are the coefficients on price and perceived discount respectively. Here, we specifically choose for the uniform distribution, as heavier tailed distributions might result in a simulated dataset in which some products do not have any purchases. For the covariance matrix of the error terms, we draw a J x J matrix from an Inverted Wishart distribution centered at the identity matrix and 25 degrees of freedom. The choice set $\{Y_{it}\}$ is generated in accordance with utility theory with the parameters as set above. The utility model is then defined by taking the utility differences with respect to the J^{th} alternative and Σ is transformed accordingly into the covariance matrix of error differences $\tilde{\Sigma}$. The hyperparameters for the prior distribution $p(\tilde{\Sigma}) \sim IW(\lambda, \Phi)$ are set with degrees of freedom $\lambda = 50$ and scale parameter $\Phi = 50I$, and the hyperparameters for $p(\Sigma_\beta) \sim IW(\nu, \Psi)$ with degrees of freedom $\nu = 15$ and scale parameter $\Psi = 2I$.

In estimating the model parameters, we exclude each individual's last three purchase occasions from the dataset, which are later on used to evaluate the accuracy of the parameter estimates and the choice probabilities. We use the OLS estimator as starting value for β and set the starting values for β_i equal to β for all i. The covariance matrix of error differences $\tilde{\Sigma}$ is set to $M_J I_J M_J'$ where I_J is the identity matrix. Lastly, we set the utilities for the purchased products

as indicated by $\{Y_{it}\}$ equal to one, and to zero otherwise, after which we add a random element determined by $U(-0.1, 0.1)$. We then run the Gibbs sampler 15,000 times, where we allow for 5,000 burn-in runs.

4.2 Results

After the sampler completes its runs, we assess convergence by inspection of trace-plots for β and $\tilde{\Sigma}$, as well as computing the Geweke test-statistic for both parameters. Figure 1 and Figure 2 show several properties of the MCMC chain for the price coefficient for the element- and trace-restricted model respectively. Inspection of these figures show strong signs of convergence for both the element- and trace-restricted model. Formal testing by means of the Geweke test statistic confirms convergence of the parameters β_i, $\tilde{\Sigma}$, β and Σ_β for both models. The autocorrelation plot of Figure 1 and Figure 2 show that the correlation weakens rapidly, and is virtually zero after 5 and 10 lags respectively. The MCMC plots for the remaining parameters for both the element- and trace-restricted model can be found in Appendix B.

After applying thinning, we obtain the posterior point estimates, corresponding to the means of the posterior densities. The estimates, including standard errors, are reported in Table 1 for both the element-restricted model and the trace-restricted model. Comparing the estimates for $\boldsymbol{\beta}$ across both models, they are of similar size and sign, with the exception of the intercept β_2. Both models estimate the appropriate signs for the price and promotion coefficients. The uncertainty in the estimates are similar as well. The Highest Posterior Density (HPD) intervals are reported in Table 2. For both the element-restricted, and the trace-restricted model, three intervals include zero in the interval. This would suggest there is posterior support that they have no effect on the utility. Comparing $\tilde{\Sigma}$ across the models, the elements of the element-restricted model are slightly larger, but we observe about equal uncertainty in the estimates with respect to the trace-restricted model. The most important observation from these results, is that the signals as estimated by both models seem to be weaker in absolute value than expected given the distributions they were simulated from in Equation 22.

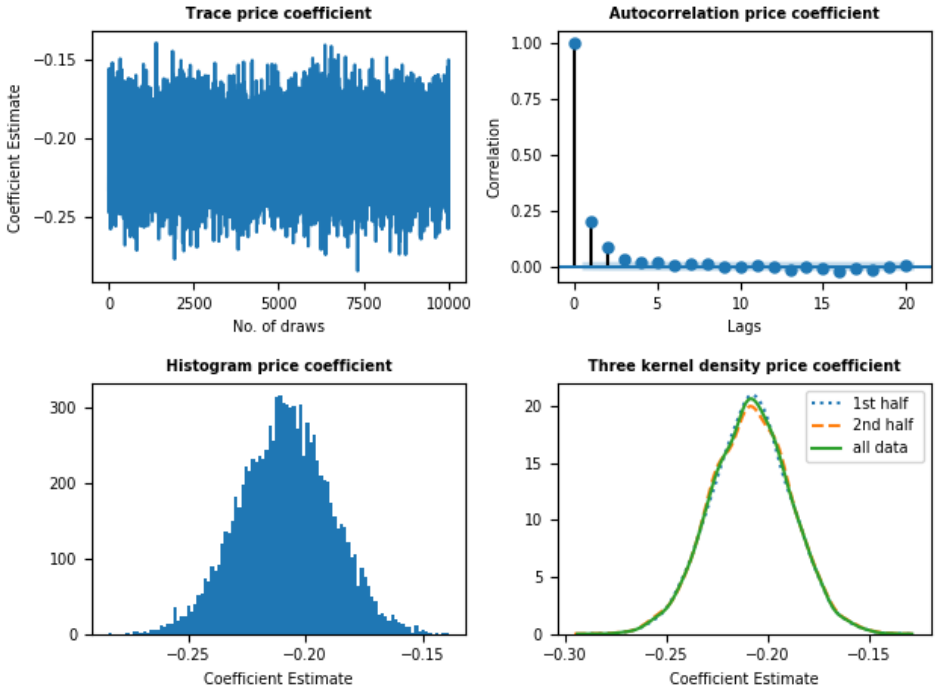


Figure 1: Convergence plots price coefficient element-restricted sampler

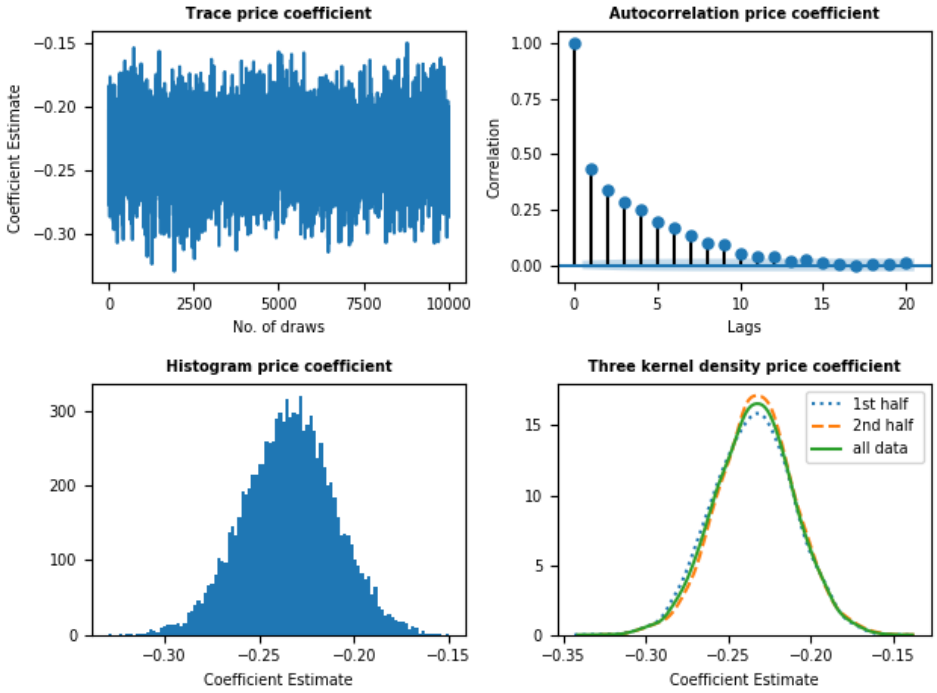


Figure 2: Convergence plots price coefficient trace-restricted sampler

.

Table 1: Posterior means of the model parameters

	Element-restricted	Trace-restricted
	Estimate (Std. Error)	Estimate (Std. Error)
Coefficients		
β_{price}	-0.209 (0.0016)	-0.235 (0.0015)
β_{promo}	0.146 (0.0031)	0.153 (0.0006)
Intercepts		
β_1	-0.144 (0.0014)	-0.047 (0.0011)
β_2	-0.041 (0.0014)	0.189 (0.0015)
β_3	0.100 (0.0015)	0.300 (0.0020)
$\tilde{\Sigma}$	$\begin{pmatrix} 1.000 & 0.053 & 0.076 \\ & (0.0018) & (0.0022) \\ & 0.979 & 0.030 \\ & (0.0036) & (0.0020) \\ & & 1.108 \\ & & (0.0040) \end{pmatrix}$	$\begin{pmatrix} 0.487 & 0.031 & 0.075 \\ (0.0019) & (0.0011) & (0.0017) \\ & 0.803 & 0.009 \\ & (0.0034) & (0.0022) \\ & & 1.710 \\ & & (0.0042) \end{pmatrix}$
Log-likelihood	- 1005.47	-951.89
WAIC	3236.56	2852.84

Table 2: Highest Posterior Density (HPD) intervals

	Element-restricted	Trace-restricted
β_{price}	(-0.245, -0.171)*	(-0.279, -0.118)*
β_{promo}	(-0.010, 0.310)	(-0.003, 0.305)*
β_1	(-0.228, -0.060)*	(-0.111, 0.021)
β_2	(-0.044, 0.130)	(-0.098, 0.278)
β_3	(0.004, 0.182)*	(0.179, 0.414)*

Note: intervals denoted with * do not include zero

Table 3: Forecast performance

	Benchmark	Element-restricted	Trace-restricted
Log predictive likelihood	-415.47	-291.77	-298.88
% correct hits	27.67	63.67	58.67
MAE	0.374	0.297	0.290

To assess overall model fit, we report the log-likelihood and the WAIC in Table 1. The trace-restricted model seems to provide a better fit tot the data, as both the log-likelihood and the WAIC are in favor of the trace-restricted model. For the trace-restricted model we obtain a WAIC of 2852.84, while for the element-restricted model we obtain a WAIC of 3236.56. In evaluating the WAIC, a smaller value represents a more favorable model. Overall, both models provide a better fit than the benchmark measure, as the log-likelihood for this model is -1335.13. This benchmark measure is calculated by using the purchase frequencies of each product during the test interval, instead of the simulated choice probabilities based on the parameter estimates. To assess predictive accuracy, we report the log predictive likelihood, the prediction hit rate and the MAE in Table 3. In terms of predictive performance, both models substantially improve on the benchmark. The benchmark log predictive likelihood is smaller than that of both models. The same can be said for the percentage of correct hits which for the benchmark is unsurprisingly equal to the purchase frequency of the most frequently purchased product. In terms of MAE, both the trace-restricted and element-restricted model improve significantly on the benchmark. When we difference the MAE_i for the benchmark with the MAE_i for both models over all individuals, we find that the means for these sets of differences is significantly different from zero. The t-statistic is 18.46 and 15.43 for the element-restricted and trace-restricted respectively. Both statistics are much larger than the critical value, allowing us to assume that both models significantly improve on the benchmark. Between the element- and trace-restricted model, the former outperforms the latter slightly. Though the MAE and the log predictive likelihood are similar across both models, the element-restricted model has a higher predictive hit rate. However, neither model is particularly accurate at 64% and 59% respectively. Here predictive hit rate is defined as the percentage of cases the highest choice probability corresponds to the product of choice.

Though predictive accuracy may be a good indicator of model fit, this study does not focus on prediction. Hence, based on overall model fit, we conclude the trace-restricted model to outperform the element-restricted model. Next, we evaluate the trace-restricted model on an empirical application.

5 Empirical application

For this research, we use data from the Dutch online retailer Picnic. Founded in 2015, the e-commerce counterpart to the conventional supermarket is only accessible via mobile applications designed for hand-held devices. Unlike conventional supermarkets, Picnic has no physical stores but delivers groceries free-of-charge to customers in over 100 cities in the Netherlands. Picnic's online store is organized along a hierarchical product tree consisting of 4 tiers, with tier 1 being the most general and tier 4 the most specific. This research will focus on analyzing the effect of introducing a new product on the other products within the 3rd category. This category is made up of the same products, but may vary in brand, volume and flavor. We believe that we can safely assume that the majority of the cannibalization will occur between products within this 3rd tier, as is in line with previous research (Srinivasan et al., 2005); (Buday, 1989). That being said, some extent of the cannibalization will be missed, as cross-category cannibalization could also occur (Van Heerde et al., 2010). However, identifying which other categories and products may be affected is outside of the scope of this research.

5.1 Data

The focus of this paper is the introduction of Zwarte Reus laundry detergent. For this product, we find a time frame for which the corresponding product category does not experience other introductions or removals. This way, we exclude possible confounding effects from other entries and exits. For these product introductions, and the incumbent products within the same category, we use daily sales data, shelf-price and perceived discount as explanatory variables. It is important to note that we exclude data from the first two weeks following the introduction of the new product. This is because the first couple of weeks surrounding a new product introduction is often characterized by short-run fluctuations in sales. For some product introductions, the additional promotional activity will catapult the number of sales in the first couple of weeks due to trial-buyers, while for other introductions a burn-in period is required to reach their full sales potential. Since we are interested in customers' change in *base behavior*, we exclude the first two weeks of sales data after the introduction of the product to ensure that the short-run fluctuations have settled.

We only consider customers that have made a purchase in both the pre- and post-introduction phase. Moreover, we only include customers who have made purchases in this category on more than 5 separate occasions, both pre- and post-introduction. We do this for the following reasons. First, due to capacity constraints, Picnic employs a waitinglist to only moderately add new cus-

tomers to their customer base. Because we do not want to include customers that have only become eligible to order from Picnic mid-way through our test phase, we exclude new customers from the dataset. Second, having more observations per customer ensures that the customer-specific estimates are more reliable. Given that the time the sampler needs to complete its runs grows with the number of customers and the number of purchase occurrences, we must be selective of how many observations to include. As such, we do not want to unnecessarily slow down the sampler by including customers whose estimates are unlikely to be very telling.

5.2 Descriptive Statistics

For the laundry detergent category, the data set consists of 5,205 trips from 767 customers over a time span of 63 weeks, excluding the two weeks directly after the introduction of Zwarte Reus. Table 4 gives an overview of the market share, selling price, and perceived discount for each brand, pre- and post-introduction.

Table 4: Data characteristics of the laundry detergent category

	Pre-Introduction			Post-Introduction		
	Market share (%)	Average price (€)	Perceived discount (%)	Market share (%)	Average price (€)	Perceived discount (%)
G'woon	78.03	1.91	0.00	77.68	1.99	0.00
Robijn	15.16	5.78	69.86	14.31	6.58	40.72
Fleuril	6.81	5.99	21.56	7.78	5.96	31.02
Zwarte Reus	-	-	-	2.13	6.59	6.38

The dataset includes three A-brands, and one Private Label brand (G'woon) which is the market-leader both pre- and post-introduction, with around 78% volume-based market share in both phases, despite any promotional activity. This is not surprising given that G'woon is significantly cheaper than the A-brands, which all operate in a similar price range. The average price is defined as the mean shelf price over the trips occurring during the pre- or post-introduction phase. Perceived discount is an indicator of promotional activity that is defined as the percentage discount a customer receives at the time of purchasing. The perceived discount is reported in Table 4 as the percentage of time that the product has a price promotion. For Robijn, Fleuril and Zwarte Reus, these perceived discounts ranged from 10% to 50% off. To avoid multi-collinearity, the perceived discount is not included in the shelf price.

In terms of switching behavior, customers tend to purchase the same brand on subsequent purchase occasions, as is exemplified by the diagonals of the switching matrices in Table 5. In fact,

89% of subsequent trips consist of repeat purchases. This is confirmed by the distribution of choice sets, as 74% and 79% of customers in the pre-introduction and post-introduction phase respectively, have only one product in their choice set. However, we cannot speak of general brand loyalty as we do not account for the marketing-mix variables. For instance, G'woon is generally very cheap, and Robijn has the most discounts which could account for what naively would be misconstrued as brand-loyalty. Both G'woon and Robijn lose some market share, after the introduction of Zwarte Reus, while the market share for Fleuril increases somewhat. However, the changes are small.

Table 5: Switching matrix laundry detergent purchases[1]

	Pre-Introduction			Post-Introduction			
	G'woon	Robijn	Fleuril	G'woon	Robijn	Fleuril	Zwarte Reus
G'woon	0.95	0.03	0.02	0.96	0.02	0.02	0.00
Robijn	0.11	0.81	0.08	0.09	0.79	0.11	0.00
Fleuril	0.20	0.23	0.57	0.15	0.21	0.63	0.02
Zwarte Reus	-	-	-	0.33	0.33	0.17	0.17

5.3 Results

Unfortunately, we cannot maintain the same procedure as executed during the simulation study. The Gibbs sampler based on the mixed accept-reject algorithm used to sample the latent utilities as discussed in Section 3.2 falters every so many runs, making it infeasible to complete sufficient runs to obtain accurate estimates. The sampler falters because the mean of the truncated normal distribution falls too far outside the bound of the distribution, making it impossible to obtain a draw that is accepted by the accept-reject method. This can be caused by the fact that some customers choose products that go entirely against the predicted choice, causing the residuals to be very large. Given that the mean used in the truncated normal distribution modeling the latent utilities is dependent on Σ which is calculated using these residuals, the mean becomes too large. To overcome this issue we alter the truncated normal distribution slightly by sampling from the following distribution instead:

$$\begin{aligned} &TN(\mu_{ijt}, \tau^2_{ijt}) \\ &\mu_{ijt} = x'_{ijt}\beta_i \\ &\tau^2_{ijt} = 1/\tilde{\sigma}_{jj} \end{aligned} \tag{23}$$

[1] The $\{i, j\}$ element of the switching matrix signifies the relative proportion of buyers that after purchasing product i continue to buy product j on a subsequent trip.

In the original algorithm we use a different mean, namely $\mu_{ijt} = x'_{ijt}\beta_i + F'(\tilde{U}_{it,(-j)} - X_{i(-j)}\beta_i)$, where F is defined as $-\sigma_{jj} * \gamma_{j,(-j)}$. Here $\gamma_{j,(-j)}$ refers to the j^{th} row of $\tilde{\Sigma}^{-1}$, with the j^{th} element removed. Furthermore, we use the inverse CDF technique (also discussed in Section 3.2) to sample the latent utilities from the truncated normal distribution, instead of the mixed accept-reject region. We realize that the change in the distribution has as consequence that the separate $J-1$ univariate distributions do not approximate the multivariate truncated normal distribution for the utilities $\tilde{U}_{it}$, as proposed by McCulloch and Rossi (1994). We also realize that this in turn may lead to biased estimates for the model parameters. Whilst recognizing the limitations of the data, we feel that it is important to publish the results here, as these results showcase the potential for this novel approach to estimating cannibalization.

To estimate the change in market structure by the introduction of Zwarte Reus, we estimate the trace-restricted MNP model separately on the pre-introduction and the post-introduction observations. The resulting parameter estimates are then used to simulate choice probabilities for both phases of the test frame using the GHK simulator. In both the pre-introduction and post-introduction phases, Fleuril is used as the base brand. Posterior point estimates, including 95% intervals, are presented in Table 6, and are based on 20,000 draws after burn-in. The additional runs with respect to the simulation study are introduced because the altered specification of the truncated normal induces less variety in the draws for the latent utilities. Furthermore, We use the same prior specification as in the simulation study ($p(\tilde{\Sigma}) \sim IW(50, 50I)$ and $p(\Sigma_\beta) \sim IW(15, 2I)$).

Similar to the simulation study, we assess convergence by means of convergence plots and more formally by evaluating the Geweke test statistic. Plots for the pre- and post-introduction estimates for β_{price} as estimated by the trace-restricted model are shown in Figure 3 and Figure 4, respectively. Inspection of the figures makes evident that the Gibbs sampler does not converge as nicely for the empirical data as it does in the simulation study. Nonetheless, formal testing points out that the parameters β_i, $\tilde{\Sigma}$, β and Σ_β do in fact reach convergence. Problematic is that the autocorrelation for the price coefficient has barely weakened, even after 40 lags. Each draw thus remains strongly correlated and thinning barely has an effect. The high autocorrelation is most likely caused by the lack of random variation in the distribution for the latent utilities. If we compare the specification in Equation 23 with the original specification, we see that the mean in Equation 23 is dependent on only one random variable, namely β_i. In the original specification, the mean is dependent on draws for β_i, $U_{it,(-j)}$ and Σ (through F), allowing for more variation

in the mean of the distribution. Given that the sampler aims to approximate $\tilde{U}_{ijt}$ by $\tilde{X}'_{ijt}\beta_i$, the lack in variation in $\tilde{U}_{ijt}$, causes β_i, and thus β to remain highly correlated.

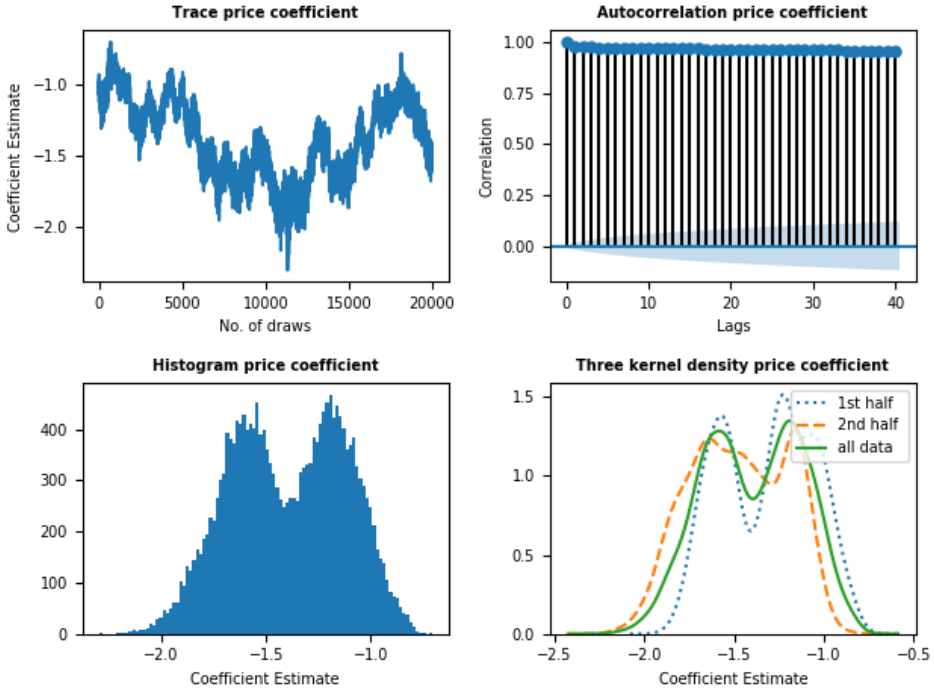


Figure 3: Convergence plots price coefficient pre-introduction

.

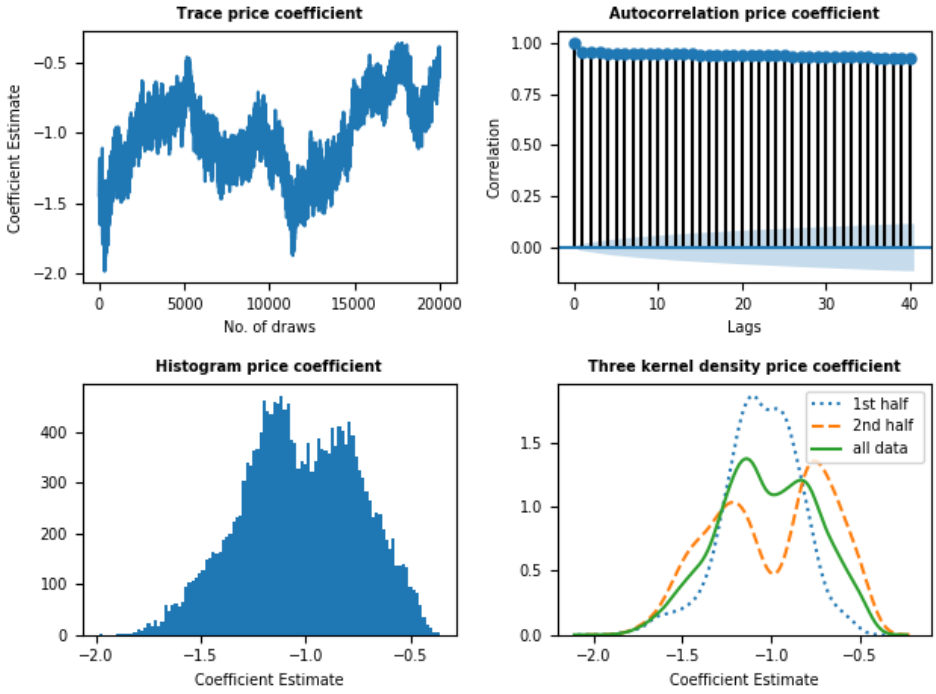


Figure 4: Convergence plots price coefficient post-introduction

.

Furthermore, the bottom two panels of both figures clearly show a bimodal distribution, more strongly so pre-introduction, than post-introduction. Bimodality may be a sign of a poor identifying restriction (Burgette & Nordheim, 2012), or that the Geweke test-statistic has falsely assured convergence (Cowles, 2002). In both cases, the posterior estimates may be biased. Given that our posterior estimate is the mean over all draws, our posterior estimate will fall right in between the two peaks containing the most posterior mass. The trace plots of the remaining coefficients are reported in Appendix B. Many, but not all, show this bimodality pattern.

5.3.1 Parameter Estimates

The parameter estimates are reported in Table 6. The coefficients β_{price} and β_{promo} have the expected signs both pre- and post-introduction. The size of the price coefficient is also in line with the price elasticity reported in the literature, while the size of the discount coefficient is larger than expected. Interestingly, intrinsic brand preference, measured by the intercepts, is mostly negative. Especially the A-brands exhibit deeply negative intercepts. This is in line however, with a growing body of literature, substantiating the theory that consumers are increasingly perceiving private label products as high-quality products similar to A-brands, at a lower price (Baltas & Argouslidis, 2007); (De Wulf, Odekerken-Schröder, Goedertier, & Van Ossel, 2005); (Erdem, Zhao, & Valenzuela, 2004). When comparing pre- and post-introduction parameter estimates, we observe that most signals become weaker in absolute sense, with the exception of the parameter for discount. Interestingly, the intercept for the private label brand G'woon switches signs, becoming positive post-introduction. The preference for the incumbent A-brand Robijn also becomes less negative. Overall, the standard errors for the coefficients and intercepts are of appropriate size. In Table 7, we report the HPD intervals for the the parameter estimates. All intervals across both phases, except for the G'woon interval, exclude zero. This means that there is posterior support that these parameters have an effect on the utility experienced by the customer, and thus have an effect on product choice. The change of sign for the G'woon intercept is not completely unexpected given that the HPD intervals for the G'woon intercept includes zero prior to the introduction of Zwarte Reus, meaning that even pre-introduction there is some posterior support for a positive G'woon intercept.

.

Table 6: Posterior means of the model parameters pre- and post-introduction Zwarte Reus

	Pre-Introduction	Post-Introduction
	Estimate (Std. Error)	Estimate (Std. Error)
Coefficients		
β_{price}	-1.397 (0.009)	-1.025 (0.029)
β_{promo}	4.407 (0.018)	5.314 (0.001)
Intercepts		
$\beta_{\text{Zwarte Reus}}$	-	-2.302 (0.019)
β_{Robijn}	-3.656 (0.015)	-1.554 (0.023)
$\beta_{\text{G'woon}}$	-1.841 (0.037)	0.927 (0.02)
$\tilde{\Sigma}$	$\begin{pmatrix} 1.273 & 0.0006 \\ (0.004) & (0.001) \\ & 0.7272 \\ & (0.004) \end{pmatrix}$	$\begin{pmatrix} 1.083 & 0.001 & -0.001 \\ (0.005) & (0.001) & (0.001) \\ & 0.956 & 0.001 \\ & (0.004) & (0.001) \\ & & 0.961 \\ & & (0.005) \end{pmatrix}$
Σ_β[2]	$\begin{pmatrix} 12.873 & -0.024 & 2.468 & -8.042 \\ (0.098) & (0.111) & (0.027) & (0.079) \\ & 1.125 & 0.128 & 0.240 \\ & (0.032) & (0.023) & (0.079) \\ & & 0.790 & -1.691 \\ & & (0.010) & (0.019) \\ & & & 5.804 \\ & & & (0.087) \end{pmatrix}$	$\begin{pmatrix} 0.625 & 0.426 & 0.106 & -0.162 & 0.058 \\ (0.017) & (0.038) & (0.015) & (0.015) & (0.008) \\ & 5.600 & 0.877 & 1.178 & 0.126 \\ & (0.056) & (0.040) & (0.024) & (0.030) \\ & & 0.800 & 0.014 & 0.065 \\ & & (0.018) & (0.015) & (0.009) \\ & & & 1.306 & -0.112 \\ & & & (0.013) & (0.012) \\ & & & & 0.413 \\ & & & & (0.008) \end{pmatrix}$
Log-likelihood	-1716.88	-2150.08

.

Table 7: Highest Posterior Density (HPD) intervals

	Pre-introduction	Post-introduction
β_{price}	(-1.881, -0.936)*	(-1.528, -0.503)*
β_{promo}	(4.290, 6.344)*	(3.250, 5.557)*
$\beta_{\text{Zwarte Reus}}$	-	(-3.656, -1.333)*
β_{Robijn}	(-4.588, -2.787)*	(-2.774, -0.008)*
$\beta_{\text{G'woon}}$	(-3.830, 0.039)	(-0.956, 2.460)

Note: intervals denoted with * do not include zero

[2]The order of variables in Σ_β is first the intercepts (Zwarte Reus, Robijn, G'woon) and then the coefficients (price, promo)

Unlike in the simulation study, we also report Σ_β modeling unobserved heterogeneity in Table 6. Similar to Allenby and Rossi (1998), we find substantial heterogeneity across consumers, especially pre-introduction. The first three diagonal elements of the matrix signify the unobserved heterogeneity in the intercepts (Zwarte Reus, Robijn, G'woon) and the latter diagonal elements report the unobserved heterogeneity in the marketing decision variables (price, promo). Pre-introduction, we observe much heterogeneity for Robijn and price, with respect to the other parameters. Post-introduction, the unobserved heterogeneity remains high for Robijn, but weakens for price, while we observe more heterogeneity for discount instead. Intuitively this means that across test phases, attitudes towards Robijn vary greatly among customers. The same can be said for price-sensitivity pre-introduction, and discount-sensitivity post-introduction. The posterior standard errors are more or less equal pre- and post-introduction and are of appropriate size.

5.3.2 Choice Probabilities

Using the estimates for β_i and Σ, we compute the choice probabilities for each customer, for each purchase occurrence, for each product using the GHK simulator. Given that we modify the sampling distribution for the latent utilities, we assume our posterior estimates are biased. To understand if and how our estimates affect the choice probabilities, we look at the same evaluation criteria as used in the simulation study but apply them differently. In the simulation study, we use % correct hits and the MAE to evaluate predictive accuracy. To do so, we withhold some of the data to create a hold-out sample. For the empirical application however, we are reluctant to withhold some of the data because it may further reduce the chance of making any cannibalization effect apparent. Furthermore, we observe a significant drop in model-fit when we reduce the minimal number of trips from 5 to 4. Thus, the % correct hits as reported on in Table 8 refers to how often we assign the highest choice probability to the chosen product in-sample. Likewise, we calculate the MAE for each customer over all T_i trips, instead of over the hold-out sample.

Despite the potential presence of bias in the estimates, the model performs well as shown in Table 8. In 94% and 95% of the cases the model correctly assigns the highest probability to the chosen product. However, this figure may be slightly flattered. This is because the customers in this dataset are quite persistent in their brand choice. As such, it is easier to predict a customer's next purchase. Furthermore, the MAE is low at 0.273 and 0.219 pre- and post-introduction, respectively. Especially, when comparing the MAE to the benchmark MAE in the parentheses. This benchmark is calculated by using the purchase frequencies of each product during the test interval instead of the simulated choice probabilities. Formal testing by means of a t-test confirms that the improvement over the benchmark is statistically significant. Here, we compute the difference between the benchmark MAE_i and the model MAE_i for all individuals in the sample, and test that the mean of these differences is significantly different from zero. Pre-introduction the t-statistic is 3.61, and post-introduction the t-statistic is 64.27. Both statistics are larger than the critical value.

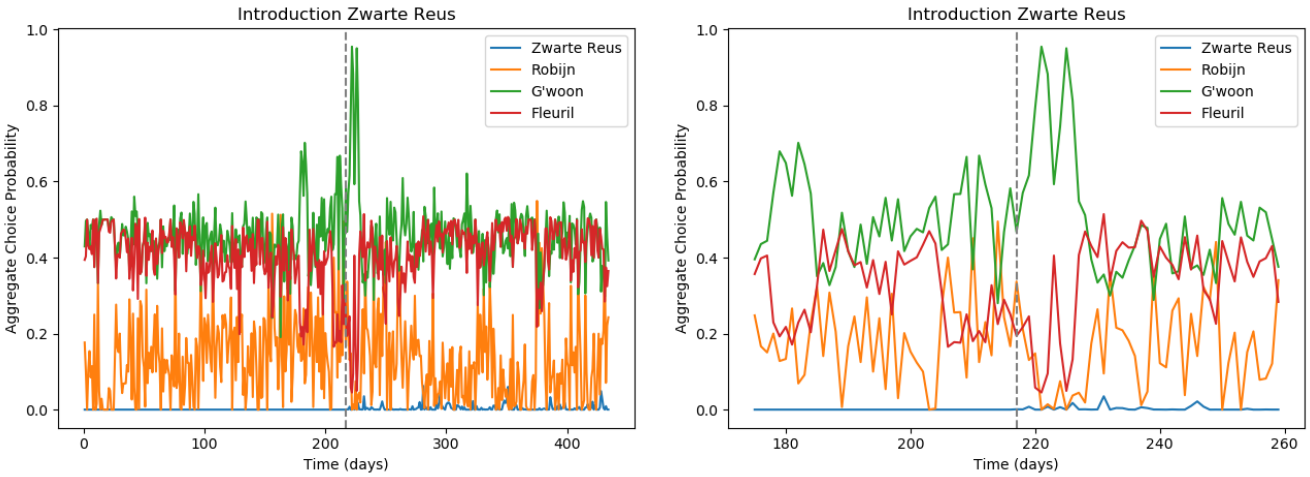
Table 8: Overall model fit pre- and post-introduction

	Pre-introduction	Post-introduction
% correct hits	94.12	94.89
MAE[3]	0.273 (0.466)	0.218 (0.482)

[3]The figure in the parentheses signifies the benchmark, where we use the purchase frequency over the testing interval instead of the simulated choice probabilities

Next we aggregate the choice probabilities over the individuals to observe whether cannibalization has occurred. This is done by averaging the choice probabilities of all individuals who made a purchase within the laundry detergent category for each point in time during the test frames. The results are reported in Figure 5. The dashed line indicates the time of introduction. Note that the dataset does not include the two weeks immediately after the introduction, to account for temporary demand shocks inherent to new product introductions. Figure 5b is a zoomed-in depiction of Figure 5a, whereby we hone in on 6 weeks prior to the introduction and 6 weeks after the introduction to improve visibility.

At first glance, the introduction of Zwarte Reus seems to have little effect, as the new product hovers around 2-3% market share and does not seem to be threatening towards the other brands in the long term. Nonetheless, the immediate peak in choice probability for the private label brand G'woon is striking. Likewise is the drop in choice probability for Robijn after the introduction. Fleuril also experiences a drop, albeit not as extreme. This is in line with earlier findings with respect to the parameter estimates. With the introduction of an additional A-brand, overall preference moves toward the private label brand, while the incumbent A-brands suffer slightly. However, we must be critical of the longevity of the effects. Given that the aggregate choice probabilities for all incumbent brands almost return to their pre-introduction levels, one could argue that the apparent effects are still part of introductory fluctuations.



(a) Actual **(b)** Zoomed-in

Figure 5: Long-term aggregate choice probabilities pre- and post-introduction Zwarte Reus, actual and zoomed-in

To observe whether including customers with fewer trips per test phase in the dataset provides a more complete view of the impact of the new product introduction, we add an additional analysis. Here, we observe shorter test-phases of 14 weeks pre- and post-introduction but include

all customers with 2 or more separate purchase occasions in both test phases. The results are reported in Figure 6. Interestingly, the pattern we observe in Figure 6 is similar to the long-term analysis. The two large peaks in G'woon choice probability, and the drop in Robijn choice probability at the beginning of the post-introduction phase closely resemble the peaks in the long-term analysis. In contrast with the long-term analysis however, we observe that after a period of 50 days post-introduction in which Robijn recovers, the aggregate choice probability for Robijn eventually stabilizes at a lower level than pre-introduction, while Fleuril stabilizes at a higher level than pre-introduction. Similar to the long-term analysis, Zwarte Reus does not become truly threatening to the incumbent brands, and we cannot speak of cannibalization. Nonetheless, Zwarte Reus does seem to have an effect on the competitive structure of the laundry detergent category.

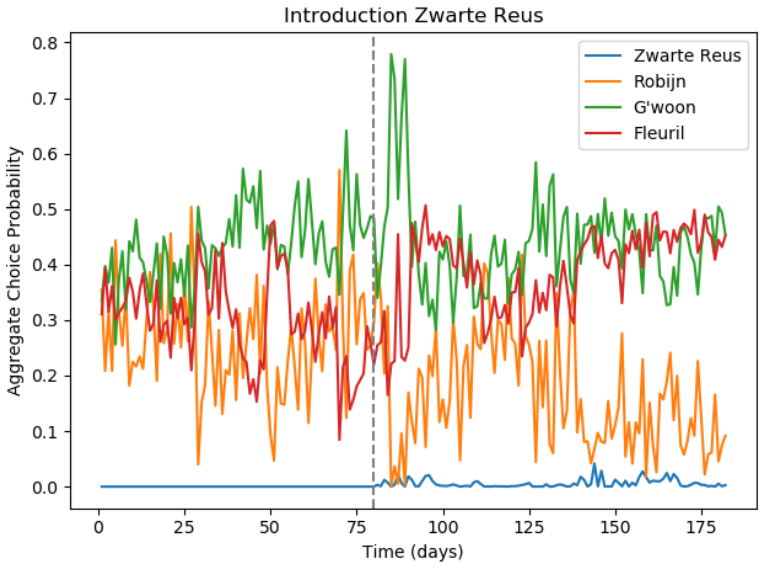
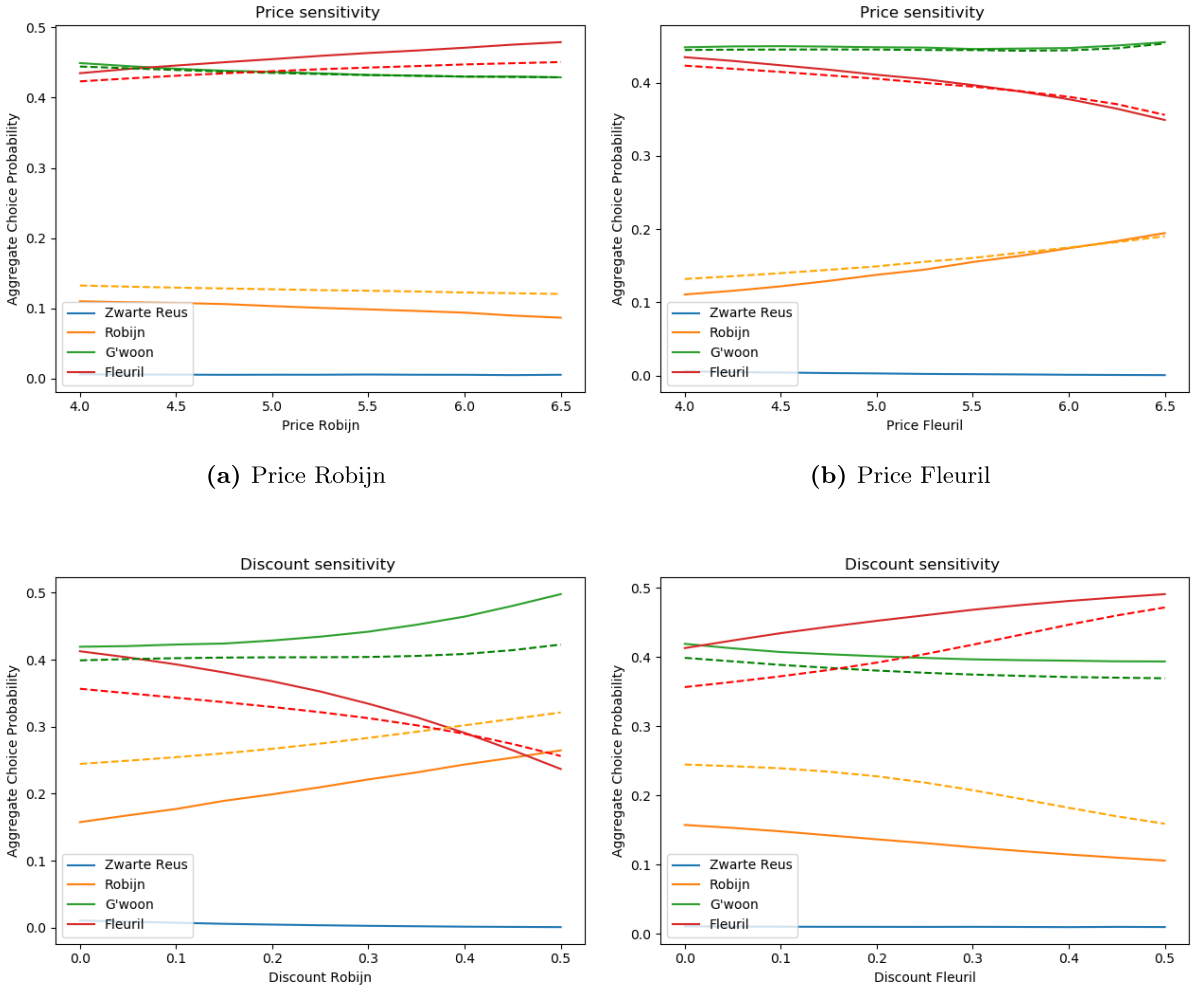


Figure 6: Short-term aggregate choice probabilities pre- and post-introduction Zwarte Reus

To analyze whether the introduction of an additional product to the category also affects individuals' sensitivity to price and promotional activity, we use the parameter estimates for β_i to model several different scenarios. First, we model the choice probabilities as a function of the price of one of the A-brands, both pre- and post-introduction. Here, we fix the price of all products to their average price, with the exception of the focal product. For this product, we range the price between 80% and 110% of its average price. Discount is set to zero for all products. Second, we model the effect of discount on the A-brands on the choice probabilities of all brands. Again, we fix the price of all products to their average price. Discount is fixed at zero for all products, except one. For this focal product, we range the perceived discount between 0% and 50%. We present the results in Figure 7.



(a) Price Robijn

(b) Price Fleuril

(c) Discount Robijn

(d) Discount Fleuril

Figure 7: Aggregate choice probabilities as a function of price and discount pre-introduction (dashed) and post-introduction (solid)

As expected, we observe that the brand experiencing the change in price loses choice share, while the competing A-brand gains choice share as the price increases. The choice share for the private label brand G'woon remains more or less unaffected. This pattern is especially apparent for a price increase in Fleuril, as reported in Figure 7b. When we model the choice probabilities as a function of the price of Robijn (Figure 7a), this effect is smaller, though still present. A potential explanation for this can be that frequent Robijn purchasers are less price-sensitive. When we compare pre- and post-introduction, we see that the post-introduction curves (solid) are slightly steeper than the pre-introduction curves (dashed). This would suggest that customers have become more price-sensitive, contradicting the parameter estimates of Table 6. However, the difference pre- and post-introduction in Figure 7a and Figure 7b is very small. Though we did not formally test it, the opposite scenario wherein customers become less price-sensitive, may lie within the margin of error.

For promotional activity, we observe that the brand with the promotion gains choice share as the promotion deepens. As expected, competing A-brands Robijn and Fleuril suffer the most when the other is on sale. Interestingly, Figure 7c shows the private label brand G'woon also benefits a little from a discount on Robijn, especially post-introduction. This is not the case when a promotion on Fleuril occurs. In general, consumers seem to be more discount-sensitive post-introduction, as we observe slightly steeper curves. With respect to brand-preference, households seem to become less favorable towards Robijn but more favorable towards Fleuril and G'woon post-introduction at their respective average prices, as made apparent by the intercepts.

Comparing the top and bottom panels of Figure 7, we clearly see that the effect for discount is more profound than that of price. Furthermore, the difference pre- and post-introduction is larger for the discount analysis than for the price analysis. This is not surprising given the parameter estimates in Table 6, where we clearly observe that the effect of discount is larger than the effect of price, in absolute sense. However, we must be careful directly comparing the parameter estimates as they operate on different units of measure. Interestingly, for the newly introduced product Zwarte Reus, none of the scenarios are attractive enough at the proposed proposition (average price, no discount). Lastly, we observe that especially the market position for Fleuril is highly influenced by price and discount. In fact, Fleuril can be market leader when offering certain discounts or when Robijn increases its price. However, the market position of Fleuril drops below that of Robijn, for high levels of discount on Robijn. The market positions for all other brands are more stable among different ranges of price and discount.

Overall, we find that the introduction of Zwarte Reus affects the sensitivities for the marketing decision variables price and perceived discount, despite the lack of quantifiable cannibalization taking place. We cannot truly speak of cannibalization because in the long term, the introduction of Zwarte Reus does not threaten the market position of the other brands. In the short term however, we do see strong changes in aggregate choice probability. As expected, the competing A-brands suffer from the introduction of an additional A-brand, but interestingly, it is the private label brand that benefits from the introduction. However, the effects are only temporary, as all brands rapidly recover to their pre-introduction choice shares.

6 Conclusion

In this paper, we set out to estimate the cannibalization effect that may occur in the context of a new product introduction. Here, cannibalization is assessed by the change in the aggregate choice probabilities of the incumbent products pre- and post-introduction. Simultaneously, we estimate the change in individuals' sensitivity to marketing decision variables such as price and perceived discount, due to new product introduction. We do so by means of the Multinomial Probit (MNP) model, a model that has gained much traction in the literature since improvements in (Bayesian) simulation methods and computing power. The most obvious advantage of the MNP model is that it allows for the modeling of unobserved heterogeneity across consumers, without being restricted by the Independence of Irrelevant Alternatives (IIA) assumption. As such, the MNP method takes preference over aggregate demand models which cannot model consumer heterogeneity, and over other discrete choice models such as the Multinomial Logit, which is bounded by the IIA assumption. In this study, we use Bayesian methods to obtain estimates for the model parameters. Though many classical methods of estimation exist as well, the use of such methods is not recommended, as they may be computationally intensive and only yield approximate results. Instead, we opt for obtaining the parameter estimates by means of an MCMC Gibbs sampler, where we sample the latent utilities alongside the model parameters, also known as data augmentation.

A well-known complication of the MNP model is the lack of identification, due to the scale- and level-invariancy of the latent utilities. Consequently, some parameters are unidentifiable, unless a restriction is placed on the covariance matrix of differences $\tilde{\Sigma}$. To do so, common practice is to set one of the diagonal elements of $\tilde{\Sigma}$ equal to one. More recently, Burgette and Nordheim (2012) have proposed an alternative identifying restriction, by fixing the trace of $\tilde{\Sigma}$ instead. Proposed advantages are that the trace restriction provides stronger identification, more interpretable results and is less prone to making extreme predictions. Moreover, fixing the trace, instead of just one element, eliminates the difficulty of determining for which product the variance is set to unit variance, which may have consequences for the parameter estimates. In this study, we compare the trace-restricted model of the aforementioned authors with the element-restricted model of Imai and Van Dyk (2004). Both studies make use of marginal data augmentation, which is said to improve the mixing behavior of the distributions, and limits the correlation between the draws by introducing and integrating out a working parameter to improve the sampler's convergence rate. Our contribution to the literature is that our model specification allows for unobserved heterogeneity across consumers by using individual-specific regression parameters, whilst not as-

suming $\mathrm{E}[\beta_i|\beta, \Sigma_\beta] \neq 0$. For neither sets of aforementioned authors, this is the case.

During a simulation study, we find that in line with the findings of Imai and Van Dyk (2004), the autocorrelation between draws weakens swiftly and is virtually non-existent after 10 lags for both models. In terms of overall model fit, we find that the trace-restricted model outperforms the element-restricted model. Though for both models, the parameter estimates seem to be slightly too small in absolute value. This has as consequence that neither model is particularly accurate in prediction. Nonetheless, based on the in-sample value of log-likelihood and the WAIC, we take preference to the trace-restricted specification.

Unfortunately, we cannot comment on the performance of the trace-restricted model on the empirical dataset for laundry detergent, provided by the online supermarket Picnic. The Gibbs sampler falters every so many runs, making it infeasible to reach sufficient runs to obtain accurate parameter estimates. We hypothesize that this is due to a lack of variation in the price and perceived discount covariates. Another potential cause is that the households included in the dataset exhibit very little switching behavior. Both contribute to the fact that the sampler has very little random variation to accomplish improvement of the parameter estimates. We find however, that if the mean of the distribution for the latent utilities is not dependent on Σ, the algorithm no longer falters. This modification is not theoretically substantiated, but does allow us to present a hypothetical use case for the trace-restricted MNP model. We realize that the parameter estimates may be inaccurate due to this modification, but continue the analysis for the sake of illustration nonetheless.

A key finding of the laundry detergent use case is that the MNP model may still detect changes in competitive structure and changes in sensitivity towards marketing decision variables, even in the case when the introduction of a new product is non-threatening to the market position of the incumbent brands. Though we observe little to no cannibalization accompanying the introduction of Zwarte Reus, we do see household preferences shift. For instance, households' stance towards the private label brand seem to become more favorable with the introduction of this additional A-brand offering. Furthermore, with the introduction of an additional product within the laundry detergent category, consumers seem to become less price- but more discount-sensitive. Such insights are easily overlooked and may be of great interest to retail managers considering to expand their product portfolios.

7 Limitations & Future research

In this research, we are limited to the data of a single, online, supermarket. Though the data contains sufficient observations, it may lack in sufficient variation in price and discount to truly get an understanding of switching behavior due to a new product introduction. The sampling methods do not falter on the simulation data, which is purposely simulated to resemble the empirical data. The only difference is that the covariates are generated with replacement, inducing more variation in covariates. Little variation in price and discount also results in less switching activity, as is apparent in our empirical data. When little switching occurs, it is more difficult to gain understanding of consumer preferences. Furthermore, if switching then does occur, it is difficult to see why this is the case, as the covariates differed very minimally. Our solution to introduce a different specification for the truncated normal distribution for the sampling of the latent utilities, forces us to question the accuracy of the resulting parameter estimates. As such, including data from other supermarkets may be needed to supplement our data and validate our findings. Alternatively, further research has to point out whether our alternative specification of the truncated normal are valid. Furthermore, it would be interesting to look at a more successful product introduction, to see whether customers truly abandon one product for another. Though the introduction of Zwarte Reus did change the competitive structure slightly, it never became threatening to the market position of the other brands. Applying our methods to a product introduction for which this is the case, would allow for a true quantification of the cannibalization effect.

In an attempt to obtain accurate individual-specific parameters, this research only included customers that have at least 5 separate purchase occasions both pre- and post-introduction. Evidently, in order to truly be able to draw conclusions regarding consumer behavior, it is necessary to have as many observations per customer as possible. But by excluding customers with fewer purchase occasions, this research put the focus on so-called 'habit' buyers. This has two consequences. First, some cannibalization goes undetected, as all purchases made by customers under the 5-trip threshold are excluded from estimation. Second, another interesting effect that often occurs in the context of new product introductions is neglected. New product introductions may entice buyers, new to the category, to start buying within the category, allowing the category to grow as a whole. Though outside the scope of this research, evaluating the effect of a new product introduction on these first-time category buyers may be of interest as well.

Furthermore, this research did not explore the online aspect of the data, which provides many

more interesting avenues for future research. For instance, product placement in an online setting, as a counterpart to the traditional display, poses an interesting topic for research. Especially when researching brand loyalty, the influence of a "Previously Purchased" page may be significant. Many possibilities for incorporating such effects exist. For instance, one could use app-event data to include this effect as a dummy variable. Alternatively, one may indirectly model the previously-purchased effect by introducing dynamics into the model. For such an approach, we suggest following Paap and Franses (2000), who include lagged utilities as an explanatory variable and alter the MNP model to a VEC specification. Such a specification would further disentangle short-term and long-term effects of the product introduction.

Another interesting avenue for further research is adding an additional layer to the Gibbs sampler. In this research we assume the individual-specific parameters β_i to follow a theoretical distribution $N(\beta, \Sigma_\beta)$. However, one could also infer the parameter β_i from data. Using features such as income, family composition, and age as explanatory variables to infer brand preference and sensitivity to marketing decision variables, could aid in obtaining an even deeper understanding of consumer purchasing behavior. For example, Nevo (2000) find that the relationship between the choice of breakfast cereal and sugar content is intensified by the age of the consumer. For this application, using app-event data to infer β_i may provide interesting insights as well.

References

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, *88*(422), 669–679.

Allenby, G. M., & Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal of econometrics*, *89*(1-2), 57–78.

Ansari, A., Bawa, K., & Ghosh, A. (1995). A nested logit model of brand choice incorporating variety-seeking and marketing-mix variables. *Marketing Letters*, *6*(3), 199–210.

Baltas, G., & Argouslidis, P. C. (2007). Consumer characteristics and demand for store brands. *International Journal of Retail & Distribution Management*.

Ben-Akiva, M. E., Lerman, S. R., & Lerman, S. R. (1985). *Discrete choice analysis: theory and application to travel demand* (Vol. 9). MIT press.

Bijmolt, T. H., Van Heerde, H. J., & Pieters, R. G. (2005). New empirical generalizations on the determinants of price elasticity. *Journal of marketing research*, *42*(2), 141–156.

Broniarczyk, S. M., & Hoyer, W. D. (2006). Retail assortment: more≠ better. In *Retailing in the 21st century* (pp. 225–238). Springer.

Buday, T. (1989). Capitalizing on brand extensions. *Journal of Consumer Marketing*, *6*(4), 27–30.

Burgette, L. F., & Nordheim, E. V. (2012). The trace restriction: An alternative identification strategy for the bayesian multinomial probit model. *Journal of Business & Economic Statistics*, *30*(3), 404–410.

Chib, S., Greenberg, E., Chen, Y., et al. (1998). Mcmc methods for fitting and comparing multinomial response models. *Economics Working Paper Archive, Econometrics*, *9802001*.

Chintagunta, P. K. (2001). Endogeneity and heterogeneity in a probit demand model: Estimation using aggregate data. *Marketing Science*, *20*(4), 442–456.

Copulsky, W. (1976). Cannibalism in the marketplace. *Journal of Marketing*, *40*(4), 103–105.

Cowles, M. K. (2002). Mcmc sampler convergence rates for hierarchical normal linear models: A simulation approach. *Statistics and computing*, *12*(4), 377–389.

Devroye, L. (1986). Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on winter simulation* (pp. 260–265).

De Wulf, K., Odekerken-Schröder, G., Goedertier, F., & Van Ossel, G. (2005). Consumer perceptions of store brands versus national brands. *Journal of Consumer marketing*.

Erdem, T., Zhao, Y., & Valenzuela, A. (2004). Performance of store brands: A cross-country analysis of consumer store-brand preferences, perceptions, and risk. *Journal of Marketing Research*, *41*(1), 86–100.

Fok, D., & Franses, P. H. (2001). Forecasting market shares from models for sales. *International Journal of forecasting*, *17*(1), 121–128.

Fok, D., & Franses, P. H. (2004). Analyzing the effects of a brand introduction on competitive structure using a market share attraction model. *International Journal of Research in Marketing*, *21*(2), 159–177.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing*, *24*(6), 997–1016.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*(6), 721–741.

Geweke, J. (1989). Bayesian inference in econometric models using monte carlo integration. *Econometrica: Journal of the Econometric Society*, 1317–1339.

Geweke, J., et al. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments* (Vol. 196). Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN.

Gielens, K. (2012). New products: The antidote to private label growth? *Journal of Marketing Research*, *49*(3), 408–423.

Hajivassiliou, V. A., & McFadden, D. L. (1998). The method of simulated scores for the estimation of ldv models. *Econometrica*, 863–896.

Hausman, J. A., & Wise, D. A. (1978). A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica: Journal of the econometric society*, 403–426.

Imai, K., & Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, *99*(467), 854–866.

Jiao, X., & van Dyk, D. A. (2015). A corrected and more efficient suite of mcmc samplers for the multinomal probit model. *arXiv preprint arXiv:1504.07823*.

Keane, M. (1994). The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte carlo evidence. *the Review of economics and statistics*, 648–672.

Keane, M. (1997). Current issues in discrete choice modeling. *Marketing Letters*, *8*(3), 307–322.

Kim, K., & Chhajed, D. (2000). Commonality in product design: Cost saving, valuation change and cannibalization. *European Journal of Operational Research*, *125*(3), 602–621.

Leeflang, P. S., Wittink, D. R., Wedel, M., & Naert, P. A. (2013). *Building models for marketing decisions* (Vol. 9). Springer Science & Business Media.

Mason, C. H., & Milne, G. R. (1994). An approach for identifying cannibalization within product line extensions and multi-brand strategies. *Journal of Business Research*, *31*(2-3), 163–170.

Mazumdar, T., Sivakumar, K., & Wilemon, D. (1996). Launching new products with cannibalization potential: An optimal timing framework. *Journal of marketing theory and practice*, *4*(4), 83–93.

McCulloch, R., & Rossi, P. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, *64*(1-2), 207–240.

McCulloch, R., & Rossi, P. (2000). A bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of econometrics*, *99*(1), 173–193.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica: Journal of the Econometric Society*, 995–1026.

McFadden, D., et al. (1973). Conditional logit analysis of qualitative choice behavior.

Meng, X.-L., & Van Dyk, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, *86*(2), 301–320.

Moorthy, K. S., & Png, I. P. (1992). Market segmentation, cannibalization, and the timing of product introductions. *Management science*, *38*(3), 345–359.

Nevo, A. (2000). A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of economics & management strategy*, *9*(4), 513–548.

Nobile, A. (1998). A hybrid markov chain for the bayesian analysis of the multinomial probit model. *Statistics and Computing*, *8*(3), 229–242.

Paap, R., & Franses, P. H. (2000). A dynamic multinomial probit model for brand choice with different long-run and short-run effects of marketing-mix variables. *Journal of Applied Econometrics*, *15*(6), 717–744.

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and computing*, *5*(2), 121–125.

Schuurman, N., Grasman, R., & Hamaker, E. (2016). A comparison of inverse-wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research*, *51*(2-3), 185–206.

Sheffi, Y., Hall, R., & Daganzo, C. (1982). On the estimation of the multinomial probit model. *Transportation Research Part A: General*, *16*(5-6), 447–456.

Srinivasan, S., Ramakrishnan, S., & Grasman, S. E. (2005). Incorporating cannibalization models into demand forecasting. *Marketing Intelligence & Planning*, *23*(5), 470–485.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Van Heerde, H. J., Srinivasan, S., & Dekimpe, M. G. (2010). Estimating cannibalization rates for pioneering innovations. *Marketing Science*, *29*(6), 1024–1039.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(Dec), 3571–3594.

8 Appendix

A Derivation Full Conditional Posteriors

The derivations as described below are described in detail by McCulloch and Rossi (1994). The probit model as explained in section 3.2 can be expressed as a standard multivariate regression model, meaning that the parameters can be sampled using standard normal regression results. Rewriting equation 3 to fit this standard regression, we obtain:

$$\begin{aligned} \tilde{U}_{it} \,/\, \tilde{\Sigma}^{\frac{1}{2}} &= +\tilde{X_{it}} \,/\, \tilde{\Sigma}^{\frac{1}{2}\prime}\beta_i + \tilde{\epsilon_{it}} \,/\, \tilde{\Sigma}^{\frac{1}{2}} \\ U^*_{it} &= X^{*\prime}_{it}\beta_i + \epsilon^*_{it} \end{aligned} \tag{A.1}$$

The above transformation simplifies the model by allowing for normally distributed errors with zero mean and unit variance. For this standard model, we know that the full conditional posterior of the regression parameter β is normal with mean and variance dependent on the OLS-estimator $\hat{\beta}$. The full conditional posterior for the covariance matrix of the error terms is an inverted Wishart distribution.

A.1 Conditional Distribution $\tilde{U}_{it}$

Given all other parameters, the sampling of the latent utilities U_{ijt}, ..., U_{iJt} can be interpreted as a linear regression model, whereby the full conditional posterior of each utility U_{ijt} follows a truncated univariate normal distribution. The truncation region is determined in accordance with the definition of utility theory, such that if individual i purchases article j at time t, $U_{ijt} > U_{ikt} \;\forall\; k \neq j$. Given that the the sampler operates on utility differences the truncation regions are determined as follows:

$$\tilde{U}_{ijt} \begin{cases} > \max\left(\tilde{U}_{it,(-j)}, 0\right) & \text{if } \; y_i = j \\ < \max\left(\tilde{U}_{it,(-j)}, 0\right) & \text{if } \; y_i \neq j \end{cases} \tag{A.2}$$

Here $\tilde{U}_{it,(-j)}$ signifies all utilities $\tilde{U}_{it}$ except the utility for the j^{th} alternative. The truncated normal has mean

$$\mu_{ijt} = x'_{ij}\beta_i + F'(\tilde{U}_{it,(-j)} - X_{i(-j)}\beta_i) \tag{A.3}$$

and variance:

$$\tau^2_{ijt} = 1/\sigma_{jj} \tag{A.4}$$

Here $F = -\sigma_{jj}\gamma_{j,-j}$, where σ_{jj} denotes the (j,j) element of Σ^{-1} and $\gamma_{j,-j}$ is the j^{th} row of Σ^{-1} with the j^{th} element removed.

A.2 Conditional Distribution β_i and Σ

To sample β_i, we stack over time and obtain a linear regression with conditional conjugate normal prior $\beta_i|\beta, \Sigma_\beta \sim N(\beta, 1 \text{ x } \Sigma_\beta)$. The full conditional posterior distribution of β_i is therefore normal with mean:

$$(X^{*\prime}_i X^*_i + \Sigma^{-1}_\beta)^{-1}(X^{*\prime}_i U^*_i + \Sigma^{-1}_\beta \beta) \tag{A.5}$$

and covariance matrix:

$$(X^{*\prime}_i X^*_i + \Sigma^{-1}_\beta)^{-1}. \tag{A.6}$$

Note that $X^{*\prime}_i X^*_i = \sum_{t=1}^{T_i} X'_{it}\tilde{\Sigma}^{-1}X_{it}$ and $X^{*\prime}_i U^*_i = \sum_{t=1}^{T_i} X'_{it}\tilde{\Sigma}^{-1}U_{it}$.

Conditional on the utilities and β_i, the results from the standard regression model imply that the

full conditional posterior of $\tilde{\Sigma}$ is an inverted Wishart distribution. Given the weakly informative conjugate prior, the inverted Wishart has scale parameter:

$$\Psi + \sum_{i=1}^{N} \sum_{t=1}^{T_i} (U_{it} - X'_{it}\beta_i)(U_{it} - X'_{it}\beta_i)' \tag{A.7}$$

and degrees of freedom:

$$\lambda + \sum_{i=1}^{N} T_i \tag{A.8}$$

where Ψ and λ are prior parameters.

A.3 Conditional Distributions β and Σ_β

To sample β we consider the part of the posterior density which depends on β. Given that we use a flat prior on β, the posterior density is proportional to the complete data likelihood:

$$p(\beta|\Sigma_\beta, \{\beta_i\}_{i=1}^N, \tilde{\Sigma}, \tilde{U}, Y) \propto \prod_{i=1}^{N} \exp(-\frac{1}{2}(\beta_i - \beta)'\Sigma_\beta^{-1}(\beta_i - \beta)) \tag{A.9}$$

This can be interpreted as the multivariate regression $\beta_i = 1 \text{ x } \beta + \nu_i$ with $\nu_i \sim N(0, \Sigma_\beta)$. The full conditional posterior distribution is then normal with mean $\hat{\beta} = (\sum_{i=1}^N 1)^{-}1 \sum_{i=1}^N \beta_i = \bar{\beta}$. The covariance matrix is given by $\Sigma_\beta \otimes (\sum_{i=1}^N 1)^{-}1 = \Sigma_\beta / N$.

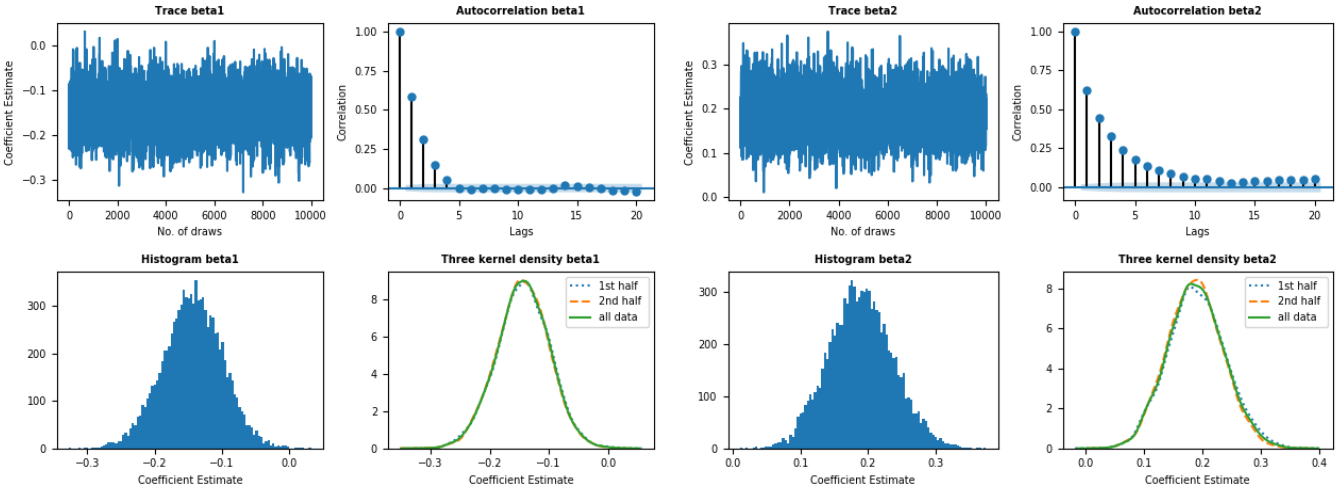
Likewise, to sample Σ_β we consider the part of the posterior density which depends on Σ_β. Given the flat prior, the posterior density is proportional to the complete data likelihood:

$$p(\Sigma_\beta|\beta, \{\beta_i\}_{i=1}^N, \tilde{\Sigma}, \tilde{U}, Y) \propto |\Sigma_\beta|^{-\frac{1}{2}(N+J+1)} \prod_{i=1}^{N} \exp(-\frac{1}{2}(\beta_i - \beta)'\Sigma_\beta^{-1}(\beta_i - \beta)) \tag{A.10}$$

The term within the exponent can also be written as $\text{tr}(\Sigma_\beta^{-1}(\beta_i - \beta)(\beta_i - \beta)')$. As such, the full conditional posterior of Σ_β is inverted Wishart with parameter $\sum_{i=1}^N (\beta_i - \beta)(\beta_i - \beta)'$ and N degrees of freedom.

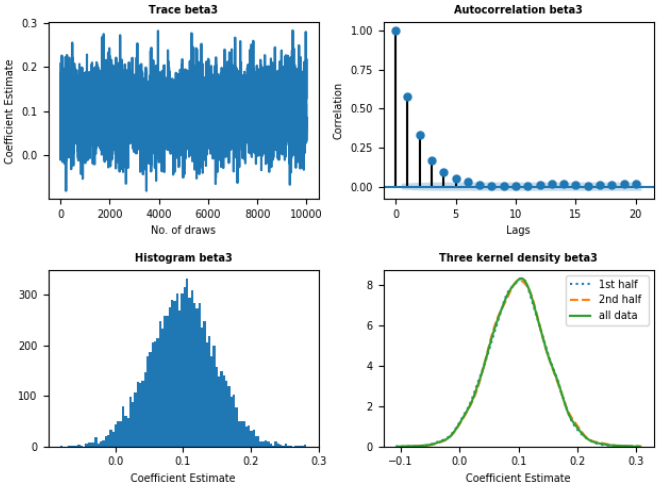
B Convergence plots

B.1 Element-restricted model

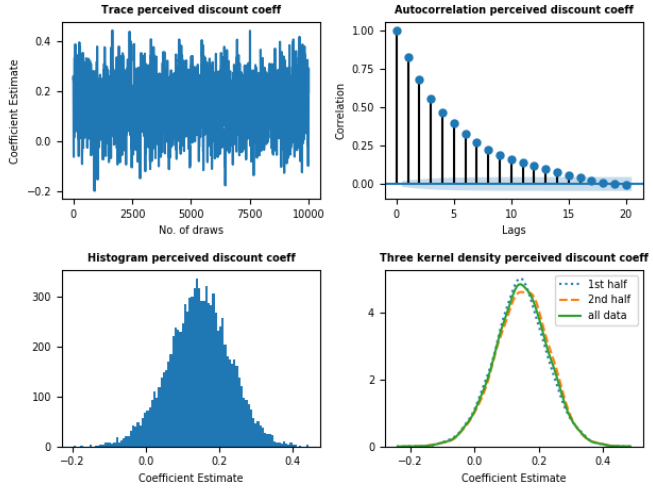


(a) β_1

(b) β_2



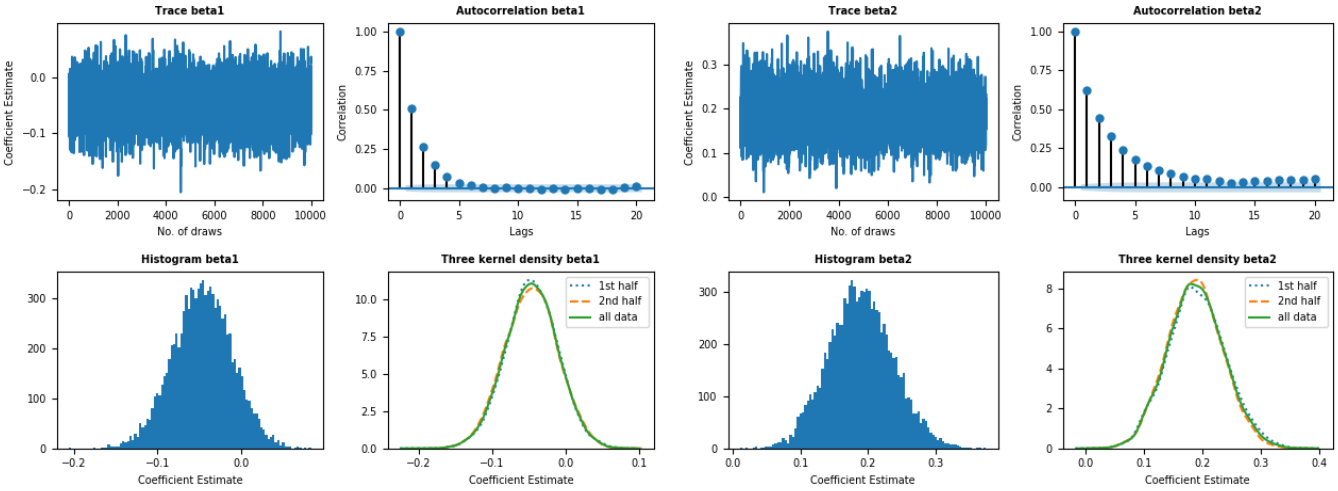
(c) β_3



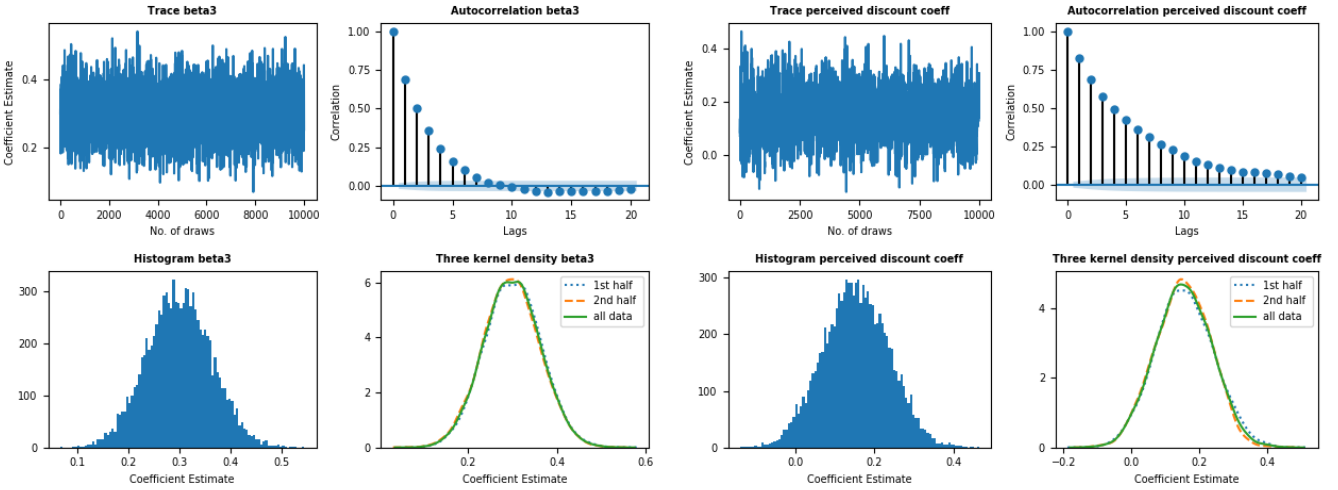
(d) Perceived Discount coefficient

Figure 8: MCMC convergence plots for coefficients for the element-restricted sampler

B.2 Trace-restricted model



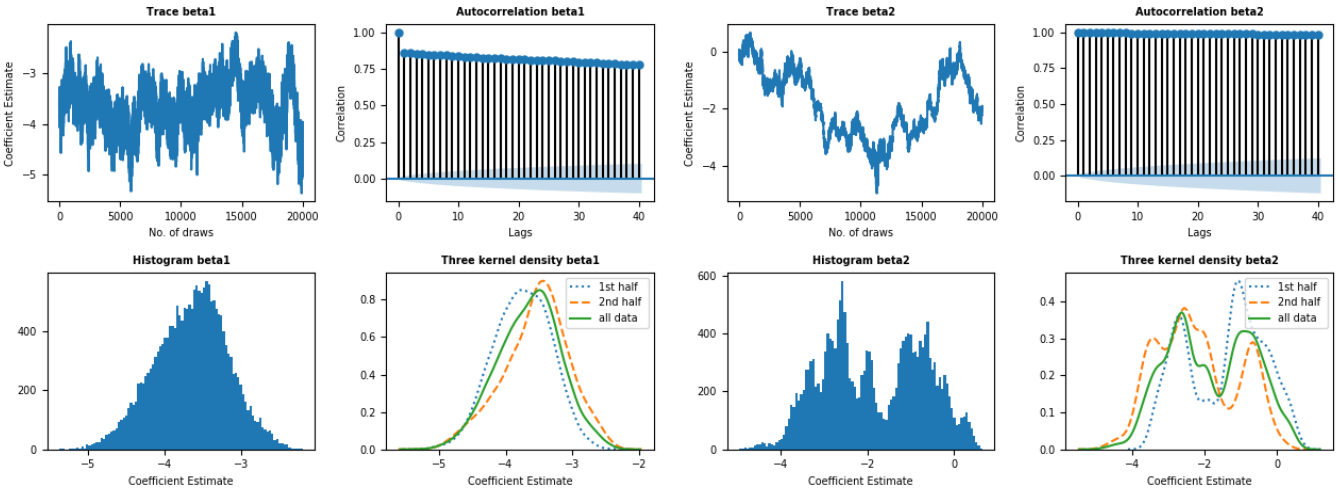
(a) β_1 **(b)** β_2



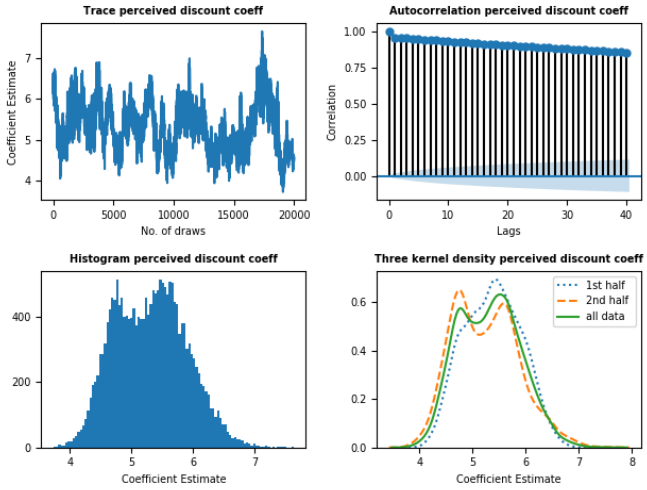
(c) β_3 **(d)** Perceived Discount coefficient

Figure 9: MCMC convergence plots for coefficients for the trace-restricted sampler

B.3 Pre-introduction (Trace-restricted model)



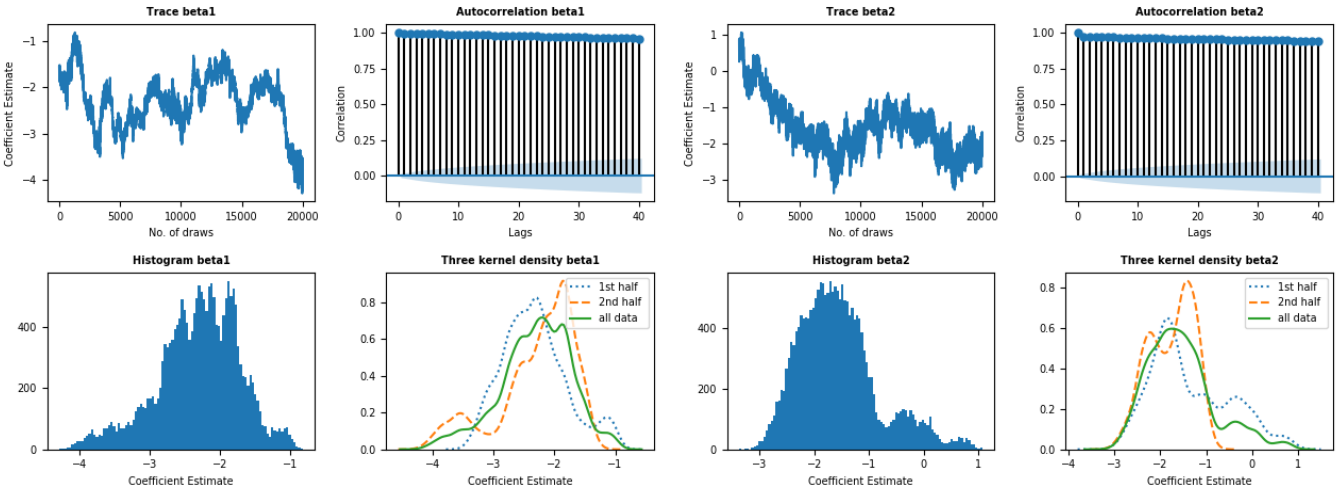
(a) β_1 **(b)** β_2



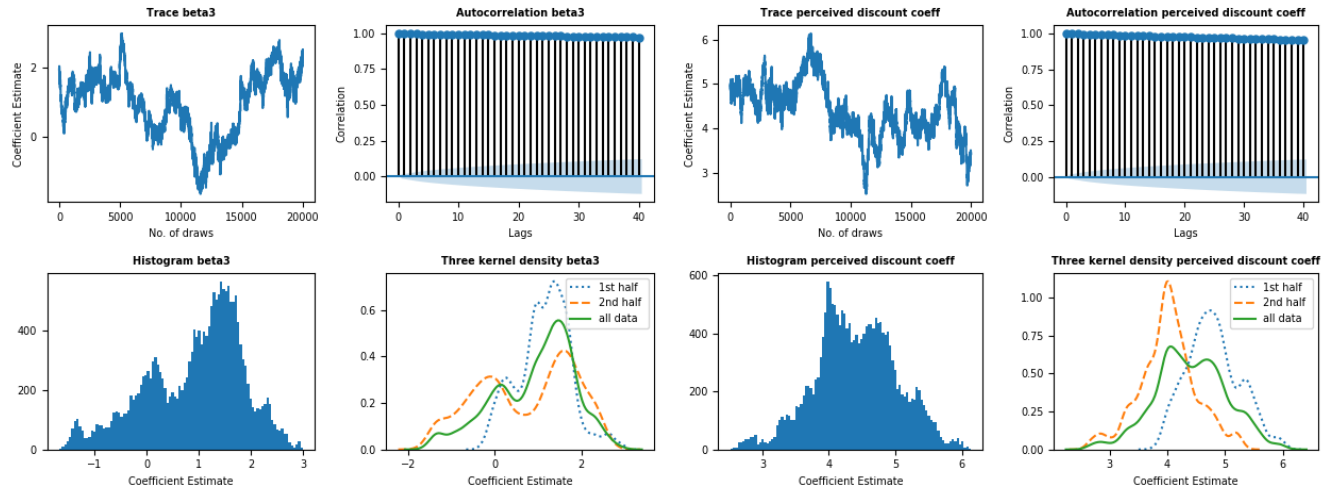
(c) Perceived Discount coefficient

Figure 10: MCMC convergence plots for coefficients pre-introduction

B.4 Post-introduction (Trace-restricted model)



(a) β_1 **(b)** β_2



(c) β_3 **(d)** Perceived Discount coefficient

Figure 11: MCMC convergence plots for coefficients post-introduction