Understanding Early Adoption of Hybrid Cars via a New Multinomial Probit Model with Multiple Network Weights

Bikram Karmakar, Department of Statistics, University of Florida,

Ohjin Kwon, Department of Marketing, Central Connecticut State University,

Gourab Mukherjee, Department of Data Sciences & Operations, University of Southern California

S. Siddarth, Department of Marketing, University of Southern California

December 5, 2021

Abstract

We analyze customers adoption of hybrid cars using automobile transaction data from the Sacramento market during the first half of 2008, a critical period when hybrid car adoption was still low. Modeling demand for durables such as cars is made more difficult by the absence of repeat purchase data and one way to address this data scarcity is to pool information across similar customers. We implement such a pooling strategy by proposing a new multinomial probit model that simultaneously accommodates different similarity structures among customers by connecting them through multiple weighted networks. Unlike the traditional multinomial spatial probit, our model links consumer connectedness to their preference and marketing mix coefficients so that each subset of the parameter vector is correlated in a unique way.

We propose and implement a novel Monte-Carlo Expectation-Maximization (MCEM) based approach to parameter estimation that significantly increases the number of consumers and choice alternatives that the model can handle. Our method modifies the computationally expensive E-step in the classical EM algorithm by a fast Gibbs sampling based evaluation. Further, it implements the M-step using a fast back-fitting method that iteratively fits weighted regressions based on associated similarity

matrices for each subset of the coefficients. We establish the convergence properties of the proposed MCEM algorithm, present computational perspectives on the scalability of the proposed method, and provide a distributed computing-based implementation.

We apply the model to sales data for compact cars from the Sacramento market. The results show that the version of our model in which the intercepts are based on the geographical closeness between consumers, and the slope coefficients on the similarity of their previously owned vehicles, fits the data the best. We explain the coefficients of the fitted model as well as present consistent estimators of the errors of the coefficients. We analyze the cross-price elasticity matrices to produce clout and vulnerability measures of the different vehicles and to produce competitive maps of the marketplace. We show how the multiple network weights explain the changes in price sensitivity of customers across geographic locations, captures the variation in brand preferences among customers, which together better estimates of a consumer's propensity to buy a hybrid.

Finally, we demonstrate how an automobile manufacturer can leverage the estimated heterogeneous spatial contiguity effects to develop more effective targeted promotions to accelerate the consumer adoption of a hybrid car.

Keywords: hybrid cars; early adoption; geographically weighted regression; multiple network weights; back-fitting; Monte-Carlo EM; multinomial probit;

1 Introduction

In order to successfully launch a product based on a new technology, it is extremely important to attract *early adopters* (Robertson, 1967, Rogers, 2010). Early adopters are thoughtful customers who play a critical role in accelerating adoption of the product so that it achieves mainstream acceptance (Moore, 1991). Because they may not be very different from the average population in their observable attributes, it is critically important to use other means to identify and target them effectively and convince them the utility of the new product is above their adoption threshold.

Automobiles are considered "durable" products, marked by lengthy inter-purchase times (the average replacement time for new vehicles is the US is about six years), relatively high prices and a high-

involvement purchase process that typically involves significant information search and processing. The current research focuses on the compact car market, circa 2007, and proposes an approach to help identify early adopters of a relatively new technology that emerged at that time, namely hybrid cars. Combining a conventional engine with a rechargeable battery to deliver more fuel efficiency, these vehicles spearheaded the push towards clean energy transportation that disrupted the traditional car market starting around the year 2003. The year 2007 represents an ideal time in the product life cycle of hybrid cars to examine consumer adoption of this new clean-energy technology and to understand how firms may potentially influence it through their marketing strategies (Heutel and Muehlegger, 2015).

Relative to consumer packaged goods (CPG), modeling consumer choice of durable goods is made more difficult by the absence of individual purchase histories, which contain valuable information about a consumer's preferences (Bucklin et al., 2008). The Random coefficient multinomial probit (RCP) is a workhorse model that has been applied to both CPG and durable goods markets to identify the impact of prices, rebates and other marketing instruments (Rossi and Allenby, 1996) on demand. Despite incorporating consumer heterogeneity, the coefficients from this model are "global", potentially constraining its ability to identify consumers who can be targeted with scarce marketing dollars. Choice models with individual parameters that are shrunk "locally" capture consumer heterogeneity in a more nuanced way, which may significantly improve predictive performance and be the foundation of a more effective targeting program. In the absence of purchase histories, marketers build local choice models by borrowing information from other consumers. Jank and Kannan (2005), Karmakar et al. (2021) report that spatial choice models that pool information based on consumer similarity deliver better predictive fit than non-spatial models. Though geographic distance is the most commonly used contiguity metric in the spatial modeling literature other proximity sources represent an opportunity to augment and improve these models. Our application explores one such measure of contiguity between buyers that is based upon the fuel efficiency of the vehicles previously owned by them.

To our knowledge, our model is the first spatial discrete choice model that (See Ch. 6 of Agresti, 2015) incorporates multiple network weights, each based on a unique consumer closeness measures, yielding a flexible model in which subsets of the parameter vector have distinctive spatially correlations. Adding these features to the model also spur the development of new statistical methods to obtain maximum likeli-

hood estimates (MLE) of the parameters. Specifically, we develop a modified Expectation-Maximization (EM) approach (Dempster et al., 1977, McLachlan and Krishnan, 2007) in which the M-step iterations incorporate backfitting. The resulting algorithm is more scalable than extant spatial choice models, greatly increasing the number of consumers and choice alternatives that the model can handle. Several desirable theoretical properties of the algorithm are demonstrated, while our empirical application highlights its practical use by making it the basis of new target marketing programs designed to more effectively accelerate consumer adoption of hybrid cars.

1.1 A Multinomial Choice Model based on multiple Network weights

The multinomial probit model (MNP), an unordered categorical data analysis method (Rossi et al., 2012), is widely used to model the discrete economic choice behavior of customers. However, parameter estimation can be computationally burdensome because the likelihood function involves hard-to-compute multi-normal orthant probabilities. A significant body of previous research, both frequentist (Keane, 1994, McFadden, 1989, Natarajan et al., 2000) and bayesian (McCulloch and Rossi, 1994, McCulloch et al., 2000), has tried to develop efficient estimation methods for the MNP (see Tosetti and Vinciotti (2019) and the reference therein for a recent review on the topic).

Spatial structures reflecting structural contiguity in choices of geographically close customers (Jank and Kannan, 2005, Yang and Allenby, 2003) have been introduced in MNP to strengthen inference in choice models with few repeated observations, as in durable products, by borrowing information across related consumers. Recently, Karmakar et al. (2021) showed that using a spatial autoregressive (SAR) structure (Ch. 6 of Anselin (2013), Ch. 2 of Banerjee et al. (2014)) on the intercepts can improve prediction of car choices. Here, we explore the advantages of having an alternative spatial structure, namely weighted regression (WR) (LeSage, 2004). While both SAR and WR are used for modeling non-stationary data, SAR allows only global shrinkage through the auto-correlation parameter whereas WR admits local regression coefficients. Thus, unlike traditional SAR, WR models are effective in the presence of spatial heterogeneity in the data (Fotheringham et al., 2003).

Furthermore, while most of the existing marketing literature has focused on a single contiguity mea-

sure based on geographic closeness, we also harness information based on other criteria such as contiguity measures based on previous vehicle choices, which recent research on automobile purchase choices reveal that features of traded-in previously owned vehicle contain significant information on new vehicle choices (Karmakar et al., 2021). However, networks with highly varying characteristics are often required to capture the customer level heterogeneity of the covariates involved in the aforementioned customer similarity measures. In these cases, using a single network weight matrix to fecilitate pooling of information across customers in the MNP may lead to incorrect inference and may deteriorate predictive performance. To accommodate possibly different similarity network structures among covariates, we consider a flexible *multiple weight* matrices based shrunken *MNP* model (`MWMNP`) that incorporates multiple network weights from different similarity measures while pooling information across customers. Recently, Fotheringham et al. (2017) developed an algorithm, abet in a Gaussian framework, for estimating parameters in a multi-scale spatial model. While spatial models with multiple weights have been used in analyzing continuous outcomes such as population (Fotheringham et al., 2017) or housing prices (Li et al., 2019), till date, to the best of our knowledge such flexible modeling approach has not been developed for discrete outcomes from a multinomial choice process. For modeling buyers car choices among different alternatives (which include hybrid choices) in the compact category we develop a modeling framework that is flexible enough to permit different components of the parameter vector to be smoothed based on different similarity networks.

1.2 Statistical Challenges

Estimating the parameters in the multiple network weights based MNP model pose several statistical challenges. To address these challenges, we develop a novel Monte-Carlo EM (MCEM) based algorithm (Wei and Tanner, 1990) for estimating the maximum likelihood estimator. We describe its key ingredients below:

(i) Monte-Carlo EM. Computationally tractable parameter estimation in the traditional MNP set-up with a substantially large sample size includes Bayesian MCMC approaches that are based on Rossi and Allenby (1996) and Roy and Hobert (2007). Frequentist maximum likelihood estimation

is usually implemented via the EM approach (Tosetti and Vinciotti, 2019). However, evaluating the expectation of the MNP log-likelihood in the E-step using the parameter estimates from the M-step involve functionals from truncated multivariate normal distributions that do not admit closed form expressions and is very computationally intensive. In our `MWMNP` model this issue is further compounded as our application case involves a large number of alternatives and significantly large sample size. We modify the MCEM approach in Natarajan et al. (2000) for evaluating expected log-likelihood to our application set-up. By bypassing direct evaluation of the likelihood function the MCEM algorithm increases scalability. It is to be noted that in these E-step calculations, our proposed MCEM algorithm fundamentally differs from the MGWR method developed for Gaussian models in Fotheringham et al. (2017). In section 4.1 we detail the development of MCEM algorithm for parameter estimation in `MWMNP`.

(ii) Backfitting algorithm based M-step. Unlike traditional MNP set-ups (sec. 3 of Natarajan et al. (2000)) or MNP with a single-weight matrix (Ch. 6 of Anselin (2013)), using an EM approach in `MWMNP` is challenging as maximization of the expected log-likelihood in the M-step is non-trivial. Inspired by the backfitting algorithm in Buja et al. (1989) that is popularly used for calibrating generalized additive models, we maximize the expected log-likelihood of our multi-weight `MWMNP`. Backfitting is a very flexible iterative algorithm which can be applied to a host of cumbersome additive models (Härdle et al., 2004, Hastie and Tibshirani, 1990). As `MWMNP` is an additive model without any closed form maximization, usage of the backfitting algorithm is a natural choice here. However, the convergence of the backfitting algorithm as well as the consistency of the solutions upon convergence is not always guaranteed (Opsomer and Ruppert, 1997), particularly once we are outside the canonical framework studied in Buja et al. (1989) and subsequently expanded in Mammen et al. (1999), Opsomer (2000), Tan and Zhang (2019). In section 4.3 we provide detailed analysis on the properties on our proposed algorithm and thereafter apply it to our application case on hybrid car adoption in Sections 5 and 6.

(iii) Large scale estimation using distributed computing. The compact car industry in the US has a large number of alternatives that have consequential market share. Analyzing consumer choices

in large (sample) observational data-sets across a wide range of discrete alternatives is computationally challenging. For scalable estimation of the `MWMNP` model in large longitudinal data-sets we harness the benefits of distributed computing. Recently, distributed computing based algorithmic developments for increased scalability and reduced computational time without sacrificing the requisite level of statistical accuracy have received significant attention. By carefully decoupling and distributing the unrelated calculations across observations in both E-step and M-step of our proposed backfitting based MCEM algorithm, we massively reduce computational complexity (see Lemma 1 and the discussions before it). To implement our proposed methodology, we develop a R-package `MWMNP` that can be freely downloaded from Github repository.

We apply the developed `MWMNP` method for analyzing car purchase data from the Sacramento, California market in 2017. We study the adoption rates for the two prevalent hybrid cars at that time in the compact category, viz, Toyota Prius and Honda Civic Hybrid . The key findings in our case-analysis are:

- By using a multiple network weights based discrete choice model, we show that a consumer's previous vehicle characteristics contain important information regarding his current vehicle choice. This information can not be captured by just considering the recent car brands sold in his/her geographic proximity. Unlike Karmakar et al. (2021), simultaneous usage of these two heterogeneous customer networks – based on previous vehicle choice and geographic proximity – is conducted here through the `MWMNP` model. This greatly increases predictive accuracy over choice models that use single weight matrix based MNP models (See Table 2 and its discussion).

- The proposed `MWMNP` model shows that for car buyers in any geographical neighborhood, the propensity for hybrid adoption increases if his/her previous vehicle had high gas mileage measured in miles-per-gallon (mpg). However, there is stark difference between the adoption probability curves for the two hybrid cars. See figure REF and its associated text in Section 5.

- Using the `MWMNP` model, we showcase how consumer sensitivity to price changes are highly heterogeneous and greatly varies across neighborhoods (See Figure 8. We found that the `MWMNP` model with geographically varying response coefficient and preference coefficients correlated based on

their previous vehicle choices is the most efficient based on predictive accuracy (see Figure 6 and its associated text for further details).

- In Section 6 we study incentive programs to increase adoptions for Toyota Prius. We consider a popular marketing method *Conquest Cash* that targets customers and provides them a fixed rebate on the cost of Toyota Prius. We show that using the proposed `MWMNP` to target consumers for the *Conquest Cash* program can create a lift of 6.8% over traditional approach.

The rest of the paper is organized as follows. In Section 2 we describe the data behind our application case. We introduce the multiple network weights based multinomial probit model in Section 3 and present a consistent, scalable algorithm for estimation the model parameters in Section 4.1. In section 4.3 we detail on the working principle and computational complexity of our estimation algorithm and demonstrate its several appealing properties. In Section 5, the results from our empirical application is presented. In Section 6, we demonstrate how the parameter estimates from the model can be used to dramatically improve the effectiveness of a promotional program that is designed to target hybrid buyers. We conclude with a discussion of the limitations of our study and some directions for future research.

2 Data

The data comes from an established market research firm that collects vehicle transaction data from dealers electronically. In this research we analyze sales in the Premium Compact category from dealers in the Sacramento market during the first six months of 2007. We are particularly interested in using the model to notionally help the marketing team of Toyota Prius, the original hybrid car, which launched in the US in the year 2000. Figure 1 shows the US market share of Prius and Civic Hybrid over years. Till 2007, hybrid cars have an increasing market share but are yet to cross the 1% market share mark. Also, Figure 1 shows that the upward trend seen in the early years flattens around 2007 which coincides with the historical decline in light vehicle retail sales in the US (Carlier, 2021). Thus, it is extremely important to study the role of efficient marketing strategies for adoption of hybrid Prius cars at this critical juncture when the market for new cars was declining in US. After seven years of launch, Prius's the main

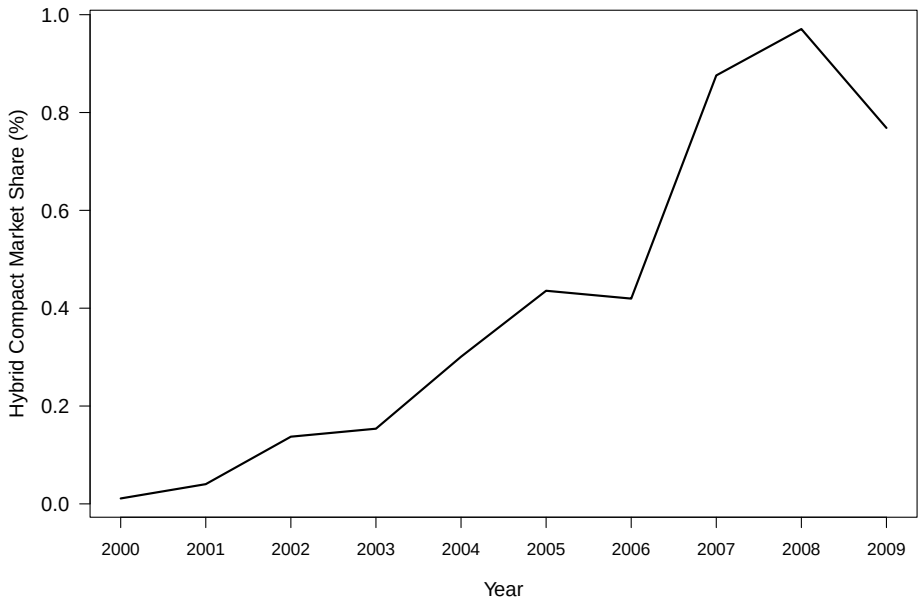

Figure 1: Market share of Prius and Civic Hybrid over the years.

marketing challenge was to accelerate adoption sufficiently by reaching beyond the innovators so that the Prius could "cross the chasm" (Moore, 2002) to mainstream acceptance in the face of direct competition from the Honda Civic hybrid, which launched in Spring 2002.

We focus on seven of the top-selling models that together account for about 64% of all sales in the Premium Compact category. The models are (a) Honda Civic, (b) Toyota Prius,(c) Toyota Corolla, (d) Nissan Sentra, (e) Honda Civic Hybrid, (f) Scion xB, and (g) BMW m3. Because we wish to examine the effect of a consumer's previously owned vehicle on their current choice decision, we restrict ourselves to only those consumers in our dataset who had also traded-in a vehicle to the dealership while acquiring a new one. The seven short-listed models accounted for 2196 transactions, which represents 70% of all Premium Compact purchases via trade-in during the first six months of 2007.The list of traded-in vehicles is much longer, including a total of 249 different models made by 38 manufacturers.

Transaction records contain information on the prices paid by each consumer, whether the vehicle was leased, financed or purchased outright, the Annual Percentage Rate (APR), down payment and monthly payment for for lease and finance contracts, manufacturer rebates and APR subvention (if any), and the residual value of the vehicle if it was leased and consumer location information.Table 1 reports key summary statistics for each vehicle model including market share, price levels and promotion spending.

Table 1: Summary statistics of the compact category Sacramento Market car sale data used in the application case

Vehicle model	**Average price($)**	**Average rebate($)**	**Average APR(%)**	**Market share(%)***
Civic	19,065	0	7.77	31.83
Prius	25,475	0	6.59	23.27
Corolla	16,895	427	7.05	19.63
Sentra	17,559	435	7.91	9.47
Scion xB	17,265	0	8.41	5.28
BMW m3	17,931	0	8.05	5.28
Civic Hybrid	23,315	0	6.69	5.24

* Market share among these seven vehicle models.

We compute Net Price as vehicle Price less rebate and the dollar value of APR subvention. Since vehicle prices are only observed for a purchased product, and models for predicting customer choices require prices for *all* alternatives, we construct the vehicle prices based on the parameters of a hedonic regression (Zettelmeyer et al., 2006), from which the Net Price can be derived. Since vehicle price is endogenous, potentially correlated with the unobservable utility error term we use the control function approach of Petrin and Train (2010), that is specifically developed for discrete choice models. The approach requires an instrumental variable, uncorrelated with consumers demand but highly correlated with the retail prices vehicle price for which, like Chintagunta et al. (2005), we use wholesale price. Thus, in the first step for each model we regress vehicle price on its wholesale price and obtain the residuals from this regression. In the second step, because the choice model is highly non-linear, following Wooldridge (2015), we augment the utility terms to include two additional variables, the residuals themselves and their interaction with vehicle price. We explicitly describe all the covariates used in the model in Section 3. Next, we describe the two different kinds of network weights that we consider in our study to incorporate different types of customer similarity characteristics.

2.1 Network Weights reflecting customer connectedness and similarity

2.1.1 Geographic location based weights ($\mathbf{W^G}$)

Spatial statistics started by recognizing that observations drawn from different locations are not independent of each other and the resulting models represented different ways of exploring the spatial dependency in observations. Therefore it is no surprise that physical location and distance play an important role in the theory and empirical application of spatial models. From a theoretical perspective, geographic contiguity is a proxy for many socio-demographic factors such as education, income, property values and wealth, which are intricately related to the value and choices of high priced purchases such as houses (Fotheringham et al., 2017) and vehicles. Empirically, in a marketing setting, Jank and Kannan (2005) find that preferences and price sensitivities in a binomial logit model of book choice are spatially correlated across geographic regions. In the automobile market, Berry et al. (1995) find truck sales are relatively high in rural areas while sedans dominate urban areas. Yang and Allenby (2003) show that a consumer preferences for a vehicle's country-of-origin (Japanese/non-Japanese) are spatially correlated based on the physical distance between consumers. Therefore geographic location is an important component of our model.

Figure 2 shows that vehicle buyers in our dataset are concentrated around Sacramento though some of them are more than 120 miles away in Sunnyvale, CA. Figure **??** and Figure 3 provide a more detailed view of two specific zipcodes, which reveal that the density of consumers varies across zipcodes, being much higher in the zipcode shown in the former than in the latter. Taken together these figures make clear that spatial distance variation may help to identify the effect of spatial contiguity on parameters. The spatial matrix based on geographical location is denoted as $\mathbf{W^G}$, with each element given as

$$\mathbf{W}^{\mathbf{G}}_{ij} = k_i \, \exp(1/d_{ij}),$$

where d_{ij} is the Euclidian geographic distance between the residential locations of consumers i and j and the constants k_i's are chosen to normalize the sum of each row to 1.

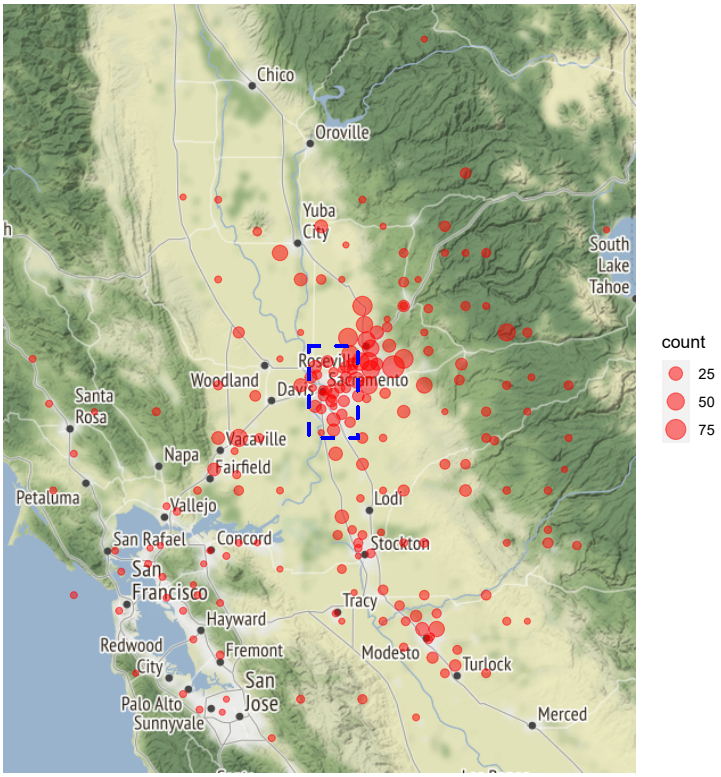

Figure 2: Geographical distribution of new compact car purchases in our data extracted from the Sacramento market.

2.1.2 Previous Vehicle similarity based weights ($\mathbf{W^V}$)

A substantial body of research has established that consumer purchasing is high in inertia and past purchases made by a consumer are highly predictive of future choices. Indeed, previous research by Karmakar et al. (2021) showed that spatial models, using very general vehicle attributes like vehicle manufacturer, country of origin, nameplate, model and the number of cylinders to measure contiguity, is highly predictive of consumer preferences for automobile models. In the current research, because the Prius was the first hybrid vehicle on the market with a extremely high fuel mpg ratings, we hypothesize that the fuel-efficiency of previously owned vehicles maybe indicative of consumer interest in this feature and thus highly relevant to modeling the spatial correlation in the demand for the Prius. Figure 3 shows the heterogeneity in mpg within and between two zip-codes in the Sacramento market. To capture customer similarity on this aspect, we constructed a contiguity matrix based on the fuel efficiency of previously owned vehicles. As a first step in this process we obtained five fuel efficiency measures from the Bureau

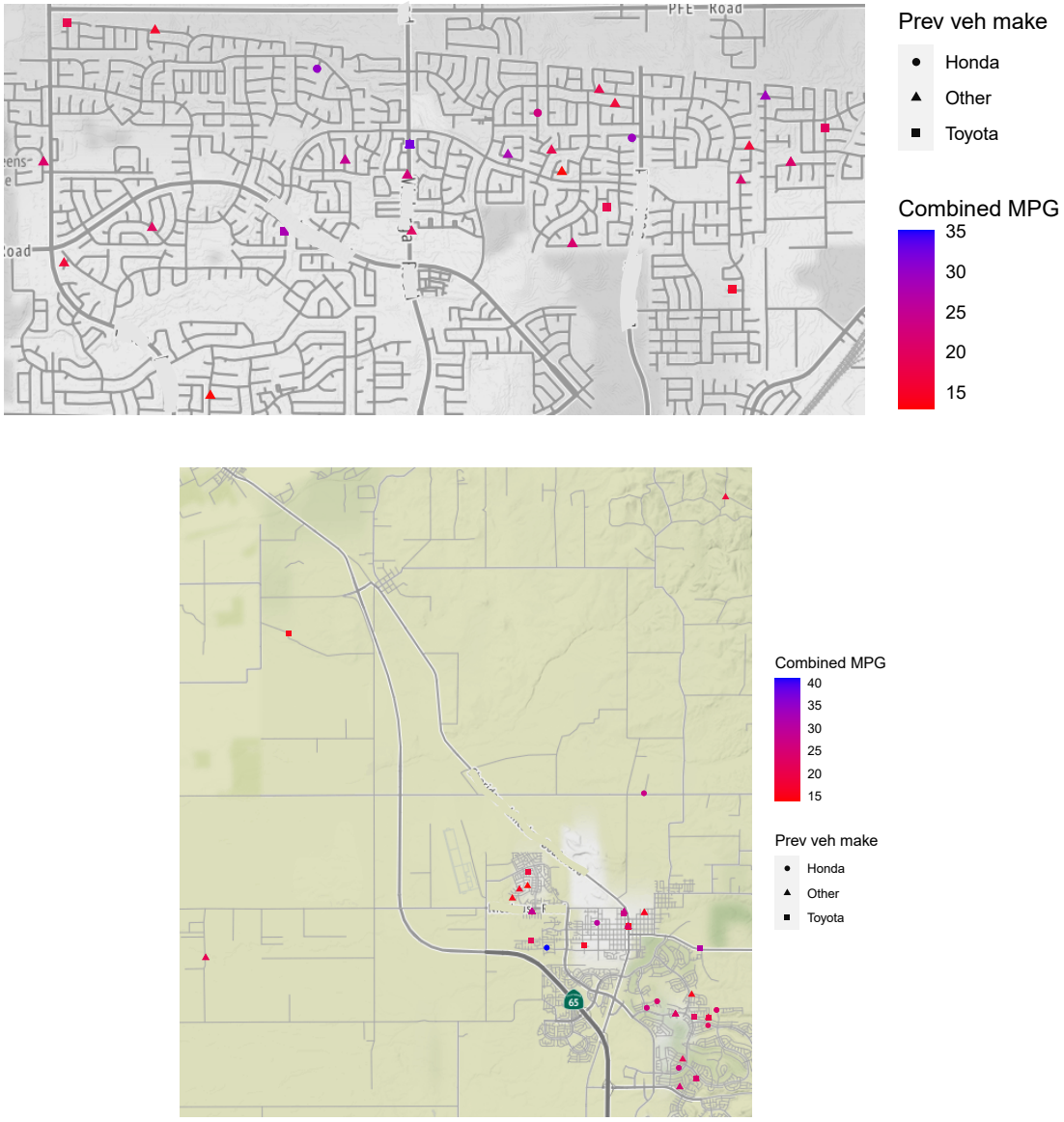

Figure 3: Two zip-codes from the Sacramento market are zoomed in to show the heterogeneity in the previous vehicle characteristics in our data set.

of Transportation for each of the 249 different traded-in models, which are are reported as the miles per gallon delivered by a car under five different conditions – city, highway, comb, UCity and UHighway.

Figure 4, a pair-wise correlation plot for these five variables shows the extremely high correlation between them. A principal component analysis (PCA) of the variance-covariance matrix yielded the following weights for each component in the principle factor: $0.359, 0.392, 0.373, 0.500$ and 0.574. Each

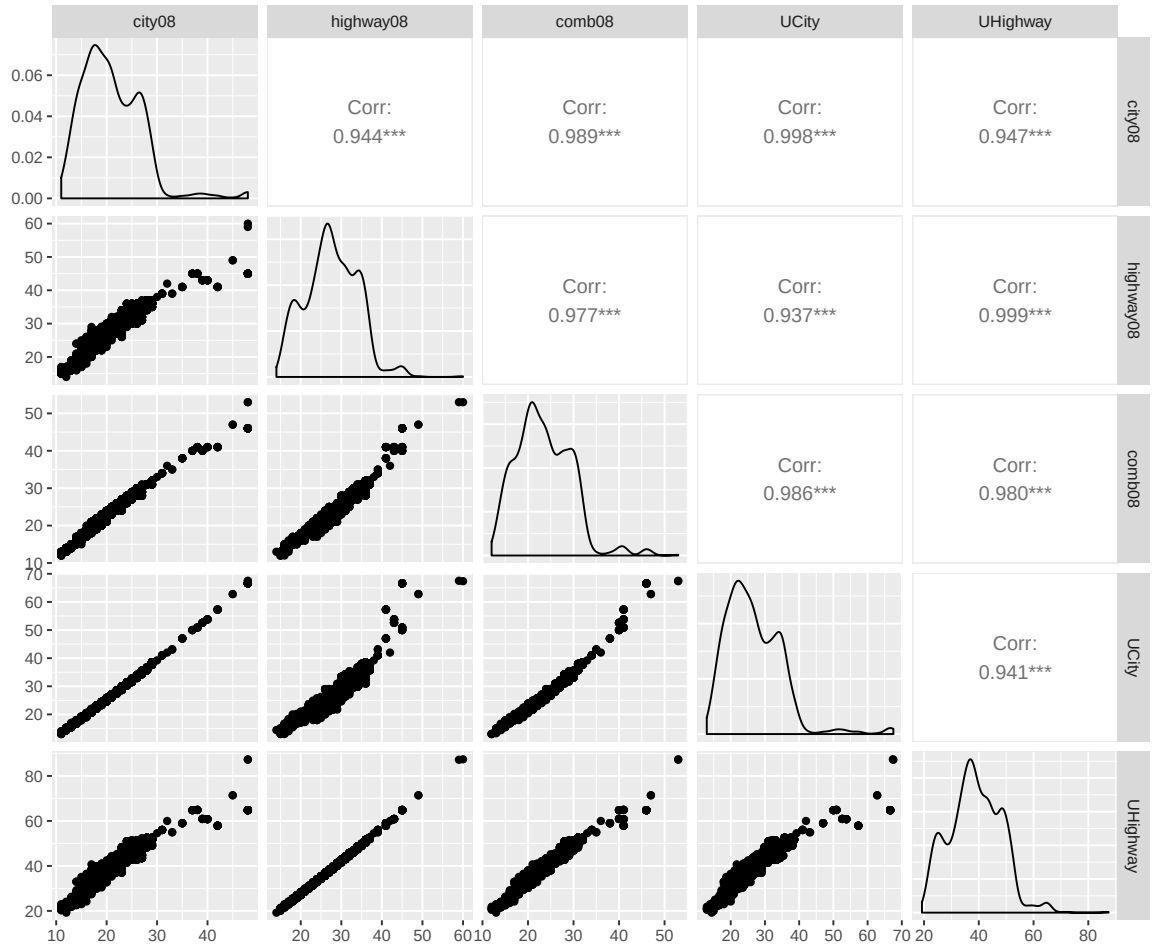

Figure 4: Pairwise plot of the five different fuel efficiency measures obtained from the Bureau of Transportation for the 249 different traded-in vehicles in our dataset.

element of the spatial contiguity matrix, W^{V} were calculated as

$$\mathrm{W}^{\mathrm{V}}_{ij} = \tilde{k}_i \, \exp(-|\mathrm{mpg}_i - \mathrm{mpg}_j|) \, ,$$

where the constants $\tilde{k}_i$'s are chosen to make the sum of elements in each row to total 1.

3 Model

Consider having C alternative car choices and O_i is the observed choice for the ith customer. Thus, $O_i \in \{1, \ldots, C\}$. Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iC})$ be the vector of utilities for the C alternatives for the consumer

i. Note that, we do not observe $\mathbf{y}_i$s but only observe:

$$O_i = \arg \max_{1 \le c \le C} y_{ic} \, . \tag{1}$$

Consider the latent utilities being generated from the following linear model:

$$y_{ic} = \alpha_{ic} + \mathbf{x}'_{ic}\beta_i + \epsilon_{ic}, \text{ for } i = 1, \ldots, N; \; c = 1, \ldots, C; \tag{2}$$

where, ϵ_{ic} have mean 0 and $\mathbf{x}_{ic}$s are p dimensional vectors corresponding to the values of the covariates for the i^{th} customer and c^{th} alternative. Stacking the c dimensional vectors $\mathbf{x}_{ij}$ as rows in the $C \times p$ matrix $X_i = [\mathbf{x}'_{i1}; \cdots ; \mathbf{x}'_{iC}]$, we rewrite (2) as

$$\mathbf{y}_i = \boldsymbol{\alpha}_i + X_i\beta_i + \boldsymbol{\epsilon}_i, \text{ for } i = 1, \ldots, N, \tag{3}$$

where, $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iC})'$. We assume that $\boldsymbol{\epsilon}_i$ are i.i.d. from Normal$(\mathbf{0}, \Sigma)$. The Gaussian regression model (3) has local coefficients. The intercepts $\boldsymbol{\alpha}_i$ which denote the preference (Rossi et al., 2012) of consumer i vary across alternatives $c = 1, \ldots, C$; the slope coefficient β_i which measures the response of customer i to changes in marketing variables is invariant across alternatives but varies across consumers. In the covariates $\boldsymbol{x}'_{ic}$ we consider the logarithm of the Net Price (defined in Section 2), and REF; also check is I have corrected replaced all n for sample size with N.

While (3) is a very flexible model, based on observing $\{O_i : i = 1, \ldots, N\}$ only, it is not estimable. We impose additional structures on the model. Powerful locally varying regression models can be constructed by using appropriate weights which reflect the underlying heterogeneity in the data (Carroll and Ruppert, 1988). Following the *Weighted Regression* (WR) model in Ch.2 of Fotheringham et al. (2003), we consider (3) with the locally linear structures on intercepts as well as the slope coefficients. Consider $N \times N$ symmetric matrices W and T with $w_{kl}, t_{kl} \ge 0$ for $1 \le k, l \le N$ and $\sum_l w_{kl}^2 = 1$, $\sum_l t_{kl}^2 = 1$ for $k = 1, \ldots, N$. For any fixed $\{\beta_i : i = 1, \ldots, N\}$, $\boldsymbol{\alpha}_i$ is the weighted least squares estimator based on

weights from W:

$$\boldsymbol{\alpha}_i = \underset{\boldsymbol{\alpha}\in\mathbb{R}^N}{\arg\min} \sum_{l=1}^{N} w_{il}^2 (\boldsymbol{y}_l - \boldsymbol{\alpha} - X_l'\beta_l)'\Sigma^{-1}(\boldsymbol{y}_l - \boldsymbol{\alpha} - X_l'\beta_l)\ . \tag{4}$$

Similarly, for fixed $\boldsymbol{\alpha}_i$s, the slopes are based on the minimizing the weighted sum of squares with weights-based on T:

$$\beta_i = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{l=1}^{N} t_{il}^2 (\boldsymbol{y}_l - \boldsymbol{\alpha}_l - X_l'\beta)'\Sigma^{-1}(\boldsymbol{y}_l - \boldsymbol{\alpha}_l - X_l'\beta), \text{ and,} \tag{5}$$

Consider the $NC \times p$ matrix $X = [X_1; \ldots; X_n]$ by appending matrices X_is by rows. Next, define $N \times N$ matrix $W_i = \text{diag}(w_{i1}, \ldots, w_{iN})$, the $NC \times NC$ non-negative definite matrix $\mathbb{W}_i$ as

$$\mathbb{W}_i(\Sigma) = (W_i \otimes I_C)\,(I_N \otimes \Sigma^{-1})\,(W_i \otimes I_C)\ , \tag{6}$$

where, $\otimes$ denotes kronecker product. Similarly, define T_i and $\mathbb{T}_i$ for $i = 1, \ldots, N$. For any fixed $A = [\boldsymbol{\alpha}_l : 1 \le l \le N]$ and $B = [\beta_l : 1 \le l \le N]$ values, define nc dimensional residual vectors $r_x(A) = [\boldsymbol{y}_1 - \boldsymbol{\alpha}_1; \ldots; \boldsymbol{y}_N - \boldsymbol{\alpha}_N]$ and $r_\alpha(B) = [\boldsymbol{y}_1 - X_1\beta_1; \ldots; \boldsymbol{y}_N - X_N\beta_N]$. Then, the solution to (4) is:

$$[\hat{\boldsymbol{\alpha}}_i | B, \Sigma] = \mathbb{P}_i(\Sigma)\, r_\alpha(B) \text{ where, } \mathbb{P}_i(\Sigma) = (Z'\,\mathbb{W}_i\,Z)^{-1}Z'\,\mathbb{W}_i \text{ and } Z = \mathbf{1}_N \otimes I_C. \tag{7}$$

Similarly, the solution to (5) is given by:

$$[\hat{\beta}_i | A, \Sigma] = \mathbb{S}_i(\Sigma)\, r_x(A) \text{ where, } \mathbb{S}_i(\Sigma) = (X'\,\mathbb{T}_i\,X)^{-1}\,X'\,\mathbb{T}_i\ . \tag{8}$$

Note that if $T_i = I_N$ and $A = 0$, then $\hat{\beta}_i$ in (5) is the generalized least squares estimator with covariance $I_N \otimes \Sigma^{-1}$. Thus, the weights t_{kl} in (5) provides a localized character to the slopes where the higher weights exerts greater influence on the MLE. In spatial applications these weights are often set based on geographic proximity between the observations, for example w_{kl} can be set inversely proportional to the distance between consumers k and l (LeSage, 2004). Often, the local characterization of

the coefficients needs calibration at different resolutions. This necessitates usage of different weights for different sets of variables. Recently, Fotheringham et al. (2017) showed that using different bandwidth scales for constructing the weight matrices W and T based on geographic distances can better explain spatial heterogeneity in the Gaussian regression model of (3)-(5). For marketing applications, as in our case, there often exists non-geodesic metrics that capture informative local patterns in the data. Karmakar et al. (2021) compared modeling product preferences using a SAR model with different kind of similarities between consumers and reported weights based on how similar consumers' previously purchased products are to each other to have more predictive power than weights based on geographical distances.

In Section 5 we show the benefits of using multiple spatial weights via (3)-(5) in modeling consumers' choices in purchasing new cars. We consider different weights W and T based on geographic as well as similarities in traded-in vehicles. It is to be noted that the Gaussian regression set-up in (3)-(5) is similar to multivariate locally weighted least squares regression (see Ruppert and Wand (1994) and the references therein) with the only difference being that the weights are not based on the local neighborhood of the concerned covariates but are based on proximity measures calculated using other variables. The MLE $(\hat{A}, \hat{B}, \hat{\Sigma})$ of (3)-(5) is usually derived by using back-fitting iterations. However, the existence of a solution depends on properties of the weights W and T. There has been extensive research on the properties of the back-fitting method for the Gaussian regression model (Mammen et al., 1999, Opsomer, 2000, Opsomer and Ruppert, 1997). In the next section, we discuss estimation in the probit set-up using back-fitting iterations in a MCEM framework.

To have interpretable models, we impose a nodal covariance structure, i.e., $\Sigma := \text{diag}(\sigma_1^2, ..., \sigma_C^2)$. The MNP model is still not identifiable (Ch. 5.2 of Train, 2009). For identifiability, we further impose the following constraints:

$$\sigma_C = 1 \text{ and } \sum_{c=1}^{C} \alpha_{ic} = 0, \quad \sum_{c=1}^{C} \boldsymbol{x}'_{ic}\beta_i = 0 \text{ for } i = 1, \ldots, N. \tag{9}$$

4 Estimation

4.1 A Monte-Carlo EM algorithm with backfitting

As we only observe O_i and not $\boldsymbol{y}_i$s, we can not directly use (4)–(5) to develop an iterative algorithm for finding the MLE as done in Gaussian regression model by Fotheringham et al. (2017). We implement an EM algorithm treating $\boldsymbol{y}_i$s as the missing data. The complete data log-likelihood based on observing the $\boldsymbol{y}_i$s is

$$-\frac{N}{2}\sum_{c=1}^{C}\log\sigma_c^2-\frac{1}{2}\sum_{i=1}^{N}\sum_{c=1}^{C}\frac{1}{\sigma_c^2}(y_{ic}-\alpha_{ic}-\mathbf{x}_{ic}'\beta_i)^2.$$

Let $\Theta=(A_{C\times n},B_{p\times n},\Sigma_{C\times C})$ represent an arbitrary parameter value. Let $\Theta^{(t)}$ be the values of the paramters at the t^{th} iteration of the EM algorithm. The E-step of the EM algorithm involves evaluating the expected log-likelihood at all possible Θ values. Note that, the unconditional distribution of the y_{ic}s is a multivariate normal based on parameters in $\Theta^{(t)}$. The expected log-likelihood for the E-step is the conditional distribution of the y_{ic}s given we had observed $\{O_i:1\le i\le n\}$ and is as follows:

$$\ell(\Theta|\Theta^{(t)})=-\frac{N}{2}\sum_{c=1}^{C}\log\sigma_c^2-\frac{1}{2}\sum_{i=1}^{N}\sum_{c=1}^{C}\sigma_c^{-2}Q_{ic}(\alpha_{ic},\beta_i,\Sigma|\Theta^{(t)})\,, \tag{10}$$

$$\text{where, } Q_{ic}(\alpha_{ic},\beta_i,\Sigma|\Theta^{(t)}):=\mathrm{E}\left\{(y_{ic}-\mathbf{x}_{ic}'\beta_i-\alpha_{ic})^2\mid O_i,\Theta^{(t)}\right\}=\{\nu_{ic}^{(t)}\}^2+(\mu_{ic}^{(t)}-\alpha_{ic}-\mathbf{x}_{ic}'\beta_i)^2$$

where, $\nu_{ic}^{(t)}=\{\mathrm{Var}(y_{ic}|O_i,\Theta^{(t)})\}^{1/2}$ and $\mu_{ic}^{(t)}=\mathrm{E}(y_{ic}|O_i,\Theta^{(t)})$ for $i=1,\ldots,N$ and $c=1,\ldots,C$. $\ell(\Theta|\Theta^{(t)})$ is subsequent maximized over Θ in the M-step. Calculating $\ell(\Theta|\Theta^{(t)})$ for any Θ is difficult as evaluating $Q_{ic}(\Theta|\Theta^{(t)})$ involves moments $\mu_{ic}^{(t)}$ and $\nu_{ic}^{(t)}$ of C-dimensional multivariate normal distributions truncated to the sets $\{\boldsymbol{y}_i:y_{iO_i}\ge y_{i1},\ldots,y_{iC}\}$. Instead of direct evaluation, we use Gibbs sampling to estimate these moments. These calculations are done in parallel for each observation i. Next, we describe the calculation of the moments $\mu_{ic}^{(t+1)}$ and $\nu_{ic}^{(t+1)}$ at the $t+1$ iterative setp of the EM algorithm based on parameter values $\alpha_{ic}^{(t)}$, $\beta_i^{(t)}$, and $\Sigma^{(t)}$ from the previous step.

E-step calculations by Gibbs sampling. By $\mathbf{y}_{i,\setminus c}$ denote the $C-1$ vector of latent utilities of consumer i except for the c^{th} alternative; similarly define $\alpha_{i,\setminus c}^{(t)}$, $X_{i,\setminus c}$ and $Z_{i,\setminus c}$ for the covariate values associated with consumer i. Set $\boldsymbol{\alpha}_i^{(0)}=\mathbf{0}$, $\beta_i=0$, $\boldsymbol{\mu}_i^{(0)}=\mathbf{0}$ for all i and variance $\Sigma^{(0)}=I_C$. For each observation i,

we conduct Gibbs sampling with an inner loop with m iterations. Starting with $\mathbf{y}_i^{(0)}$ being a random draw from normal with mean $\boldsymbol{\mu}_i^{(t)}$ and variance $\Sigma^{(t)}$, for $k = 1, \ldots, m$:

> Set $y_{ic}^{(k)} = y_{ic}^{(k-1)}$ for $c = 1, ...C$. Update $y_{ic}^{(k)}$ sequentially for $c = 1, \ldots, C$ from truncated normal with mean $\alpha_{ic}^{(t)} + \mathbf{x}'_{ic}\beta_i^{(t)} + \eta_{ic}^{(t)} \left(\mathbf{y}_{i,\backslash c}^{(k)} - \alpha_{i,\backslash c}^{(t)} - X_{i,\backslash c}\beta_i^{(t)} \right)$, and standard deviation $\sigma_c^{(t)}$; $\eta_{ic}^{(t)}$ is the regression coefficient vector which is $\mathbf{0}$, in the diagonal covariance case. If $O_i = c$, the truncation is from $\max \mathbf{y}_{i,\backslash c}^{(k)}$ to ∞. When $O_i \neq c$, the truncation is from $-\infty$ to $\max \mathbf{y}_{i,\backslash c}^{(k)}$.

Using the values $y_{ic}^{(k)}s$, after burn-in and thinning, calculate their average as $\mu_{ic}^{(t+1)}$ and standard deviation as $\nu_{ic}^{(t+1)}$. We use $\underline{m}$ values for burn in and $\overline{m}$ for calculating the $(t+1)^{\text{th}}$ iteration values. The vectors $\boldsymbol{\mu_i}^{(t+1)} = \{\mu_{ic}^{(t+1)} : c = 1, \ldots, C\}$ and $\boldsymbol{\nu_i}^{(t+1)} = \{\nu_{ic}^{(t+1)} : c = 1, \ldots, C\}$ are passed on to the M-step.

M-step: updating parameters by backfitting. At the $(t+1)^{\text{th}}$ iterative step in the EM algorithm, we maximize $\ell(\Theta|\Theta^{(t+1)})$ over Θ in the M-step. This exercise decouples into two separate maximization problems, one over A, B and the other over Σ.

We maximize $\ell(\Theta|\Theta^{(t+1)})$ over A, B by using the backfitting algorithm (Hastie and Tibshirani, 1990). The working principle behind the backfitting method is to (a) optimize the slope coefficients corresponding to each spatial weight matrix separately assuming that all the parameters for the other are known (b) cycle through the blocks of covariates corresponding to different weights estimating their parameters such that the expected likelihood is maximized (c) iterate till convergence. This again can be done in parallel for each consumer i but the parallel machines need to communicate to update the residuals. We pool the outputs across the n parallel machines used in the E-step and use $\{\boldsymbol{\mu}_i^{(t+1)} : i = 1, \ldots, N\}$ as input in the M-step. Based on the constraints in (9) we first update $\boldsymbol{\mu}_i^{(t+1)}$ as $\boldsymbol{\mu}_i^{(t+1)} - C^{-1} \sum_{c=1}^{C} \mu_{ic}^{(t+1)}$. The other inputs in the M-step are $\{X_i, \boldsymbol{\alpha}_i^{(t)}, \boldsymbol{\beta}_i^{(t)} : i = 1, \ldots, N\}$. For $i = 1, \ldots, N$, we define $\mathbb{P}_i^{(t)} = \mathbb{P}_i(\Sigma^{(t)})$ and $\mathbb{S}_i^{(t)} = \mathbb{S}_i(\Sigma^{(t)})$ using (7) and (8) respectively. With the initialization $\check{\boldsymbol{\alpha}}_i^{(0)} = \boldsymbol{\alpha}_i^{(t)}$ and $\check{\beta}_i^{(0)} = \beta_i^{(t)}$ we run an iterative inner loop. For $l = 1, \ldots, b$:

(i) Calculate partial residuals for the intercepts: $r_\alpha^{(l)} = [\boldsymbol{\mu}_1^{(t+1)} - X_1\check{\beta}_1^{(l-1)}; \ldots; \boldsymbol{\mu}_n^{(t+1)} - X_n\check{\beta}_n^{(l-1)}]$.

(ii) Update intercepts in parallel machines: for $i = 1, \ldots, N$ calculate $\check{\boldsymbol{\alpha}}_i^{(l)} = \mathbb{P}_i^{(t)} r_\alpha^{(l)}$.

(iii) Calculate partial residuals for the slope coefficients: $r_x^{(l)} = [\boldsymbol{\mu}_1^{(t+1)} - \check{\boldsymbol{\alpha}}_1^{(l)}; \ldots; \boldsymbol{\mu}_n^{(t+1)} - \check{\boldsymbol{\alpha}}_n^{(l)}]$.

(iv) Update slope coefficients in parallel: for $i = 1, \ldots, N$ calculate, $\check{\beta}_i^{(l)} = \mathbb{S}_i^{(t)} \, r_x^{(l)}$.

(v) Center the slope coefficients: $\check{\beta}_i^{(l)} = \check{\beta}_i^{(l)} - \kappa_i \mathbf{1}_p$ where, $\kappa_i = C^{-1} \mathbf{1}_C' X_i \check{\beta}_i^{(l)}$.

Set $\boldsymbol{\alpha}_i^{(t+1)} = \check{\boldsymbol{\alpha}}_i^{(b)}$ and $\beta_i^{(t+1)} = \check{\beta}_i^{(b)}$. Note that, $\mathbb{P}_i^{(t)}$ and $\mathbb{S}_i^{(t)}$ being invariant in the inner loop for the M-step greatly reduces the computational burden. As such as the weight matrices W and T are fixed through out the estimation algorithm, $\mathbb{P}_i^{(t)}$, $\mathbb{S}_i^{(t)}$ only need to be updated based on changes in the diagonal matrix $\Sigma^{(t)}$. Lastly, we update the diagonal covariance matrix parameters: for each $c \neq 1$ the standard deviations are updated as

$$\sigma_c^{(t+1)} = \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(\nu_{ic}^{(t+1)} \right)^2 + \frac{1}{N} \sum_{i=1}^{N} (\mu_{ic}^{(t)} - \alpha_{ic}^{(t+1)} - \mathbf{x}_{ic}' \beta_i^{(t+1)})^2 \right\}^{1/2}.$$

The outer loop of the MCEM algorithm is continued for increasing values of t until the parameter values convergence.

4.2 Scalability: Parallel implementation and R-package `MWMNP`

For our application case, we need the MCEM algorithm to be scalable in sample size N as well as the number of alternatives C. By using N parallel machines we greatly reduce the computational time. However, note that the N parallel machines need to communicate at the end of E-step for pooling information on the conditional expectation and variance as well as at the end of the M-step for calculating the updated covariance parameters. Figure 5 shows the schematic of the parallel architecture used in the algorithm. We optimize the computational time by using minimal updates in the E-step where the mean and variance parameter of the truncated normal distributions was evaluated only once and stored thereafter in the inner loop. Similarly, in the M-step $\mathbb{P}_i^{(t)}$ and $\mathbb{S}_i^{(t)}$ need evaluation only once. For efficient evaluation of $\mathbb{S}_i^{(t)}$, we store the singular value decomposition of X and leveraging the diagonal structure of $\mathbb{T}_i^{(t)}$ we update $\mathbb{S}_i^{(t)}$ by only $O(p^3)$ computations. Let $X = UDV$ be the singular value decomposition of X. Note that, at the $(t+1)^{\text{th}}$ outer loop iteration of the MCEM algorithm, the l^{th} inner loop iteration of the M-step involves

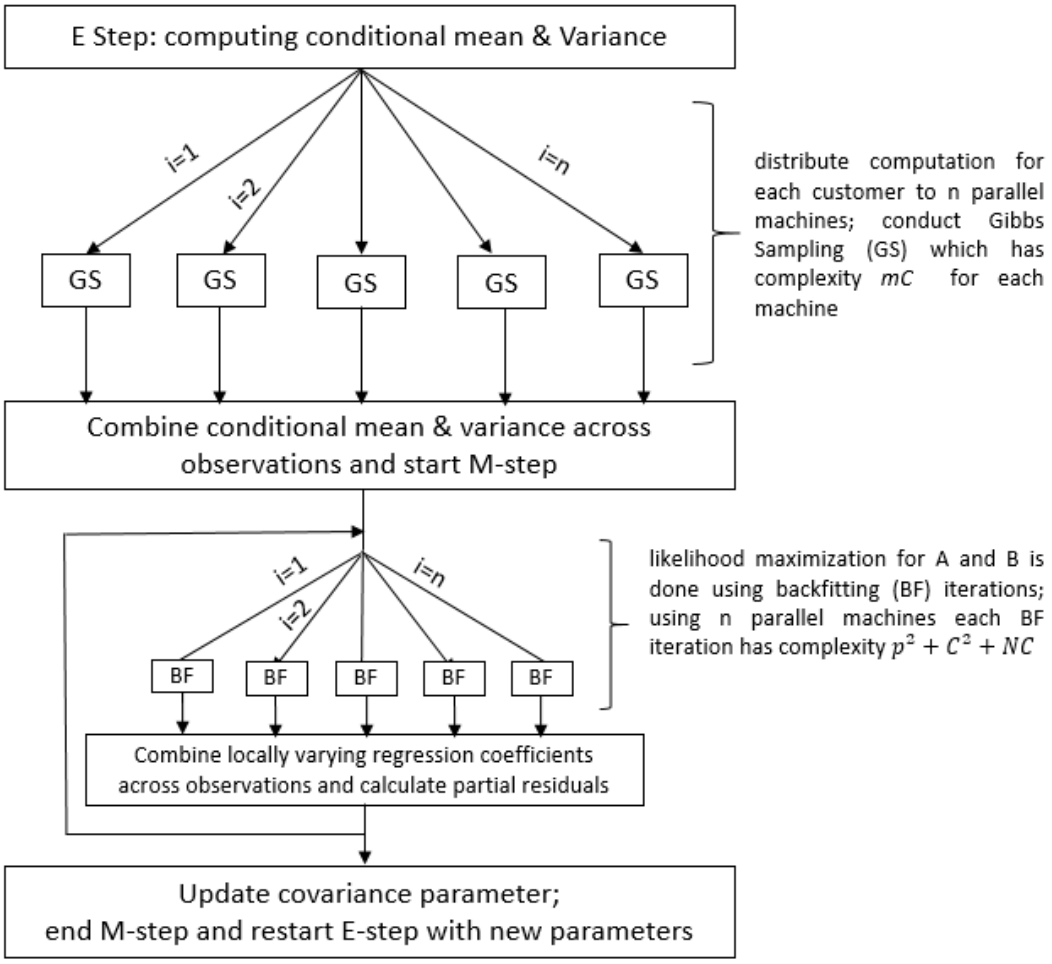

Figure 5: Schematic of MCEM algorithm used for fitting multi-weight `MWMNP` model

computing:

$$\check{\beta}_i^{(l)} = \mathbb{S}_i^{(t)}\, r_x^{(l)} = V' D^{-1} J(t,i)\, r(t,l)\ \text{, where,} \tag{11}$$

$$J(t,i) = \left(\sum_{k=1}^{NC} \tau_k(t,i)\, U'_{(k)} U_{(k)} \right)^{-1},\ r(t,l) = \sum_{k=1}^{NC} \tau_k(t,i)\, r_{x,k}^{(l)} U'_{(k)} \tag{12}$$

and $U_{(k)}$ is the k^{th} row of U; $\tau_k(t,i)$ is the k^{th} diagonal entry of $\mathbb{T}_i^{(t)}$. Computing $\{\check{\beta}_i^{(l)} : l = 1, \ldots, b\}$ knowing the SVD of X has complexity $\mathcal{O}(p^3 + b(NC + p^2))$ at the $(t+1)^{\text{th}}$ outer loop iteration. Similarly, computing $\{\check{\boldsymbol{\alpha}}_i^{(l)} : l = 1, \ldots, b\}$ is $\mathcal{O}(C^3 + b(NC + C^2))$. The following lemma documents the computation complexity of the proposed MCEM algorithm.

Lemma 1. *Distributed across n machines, the computational complexity for running τ iterations of the*

MCEM algorithm in Section 4.1 is

$$\mathcal{O}\left(NC\left(p^2+\tau\left[b+\frac{m}{N}+\kappa^3\frac{C^2}{N}+b\kappa^2\frac{C}{N}\right]\right)\right) \text{ where } \kappa=\max(1,p/C).$$

The lemma shows that for large samples ($N\to\infty$) and p fixed, the complexity of running the MCEM algorithm with stopping time τ is $\mathcal{O}\big(C(m+\tau bN)\big)$ with N parallel machines. Using distributed computing we have reduced the complexity to be linear in sample size N instead of N^3 (Fotheringham et al., 2017). However, the computation complexity here also depends on the convergence time τ of the MCEM algorithm as well as m and b. In practice, we use moderate b and large m. In Sec. 8.2 of the supplement we present simulation results on the timing and performance of the proposed algorithm along with a discussion on suitable choices of m and b. The R-package `MWMNP` is developed to implement the proposed distributed computing based MCEM algorithm.

4.3 Convergence Properties

The convergence properties of the proposed MCEM algorithm depends on the choices of m, b as well as the weights W and T. Lemma 2 shows that the Gibbs sampler used in the E-step is geometrically erdogic and so, we can throw away an initial small number ($\underline{m}$) of draws for burn-in and use the rest $\overline{m}$ draws to calculate the average.

Lemma 2. *For each $i=1,\ldots,N$ and for any fixed $\beta_i^{(t)}\in\mathbb{R}^p$, $\boldsymbol{\alpha}_i^{(t)}\in\mathbb{R}^C$ and $\Sigma_i^{(t)}\succ 0$, the Gibbs sampler for arm i at the $(t+1)^{th}$ iteration of the MCEM algorithm is geometrically ergodic.*

The lemma follows by directly checking the drift condition in Chapter 15 of Meyn and Tweedie (2012) and is proved in the supplement. For understanding the solutions from the backfitting iterations in the M-step, consider $\mathcal{P}_i^{(t)}=(I_C-C^{-1}\mathbf{1}\mathbf{1}')\mathbb{P}_i^{(t)}$ and $\mathcal{S}_i^{(t)}=(I_C-C^{-1}\mathbf{1}\mathbf{1}')X_i\mathbb{S}_i^{(t)}$. Define $\mathcal{P}^{(t)}$ and $\mathcal{S}^{(t)}$ which are square matrices of dimension NC, by combining $\{\mathcal{P}_i^{(t)}:i=1,\ldots,N\}$ and $\{\mathcal{S}_i^{(t)}:i=1,\ldots,N\}$ respectively by rows. In the existing literature, there exist strong sufficient conditions on symmetric smoothing matrices that ensure convergence of backfitting equations. However, both $\mathcal{P}^{(t)}$ and $\mathcal{S}^{(t)}$ are non-symmetric. By corollary 4.3 of Buja et al. (1989), we know that if $||\mathcal{P}^{(t)}\mathcal{S}^{(t)}||<1$ for any matrix

norm, then the backfitting iterations converge as $b \to \infty$. This constraint can be made independent of t provided the condition number of $\Sigma^{(t)}$ is bounded which is a reasonable restriction in these applications. Consider the following verifiable assumption on the weight matrices which does not need to be checked for every iteration and can be checked only once at the beginning of the algorithm. Define $\bar{W}^{(2)} = W \circ W - C^{-1}\mathbf{1}_N\mathbf{1}'_N$ where $\circ$ is the hadamard product and $J_i = e_i \otimes (I_C - C^{-1}\mathbf{1}_N\mathbf{1}'_N)$ where e_i is the unit vector in $\mathbb{R}^N$ corresponding to the ith coordinate. Also, for $i = 1, \ldots, n$ consider the non-negative definite matrices $T_i^{(2)} = X(X'T_i^2X)^{-1}X'T_i^2$ that only depends on X and T. The following assumption is on *maximum row sum norm* ($||\cdot||_\infty$) of the product of the two weight matrices. The maximum row sum norm used here has been previously used in Opsomer and Ruppert (1997) (see remark 3.1) and facilitate a simpler theoretical proof.

Assumption 1. Assume $\sup_{t\geq 1}[\sup_c \Sigma_{cc}^{(t)} / \inf_c \Sigma_{cc}^{(t)}] \leq L$, i.e., the difference between the largest and smallest diagonal entry of $\Sigma^{(t)}$ is always bounded by L and that the following condition holds:

$$\max_{i=1,\ldots,N} ||\, \bar{W}^{(2)}\, J_i\, T_i^{(2)}\, ||_\infty < L^{-1}. \tag{13}$$

The following theorem shows that under assumption 1, if we run the algorithm long enough, then the iterates at some point will get arbitrarily close to the MLE. The theorem is proved in the supplement.

Theorem 1. *Let Θ^* be a maximizer of the likelihood of observing $\{O_i : i = 1, \ldots, N\}$ in the model (1)-(5). Under assumption 1, for any $\epsilon > 0$, there exists a neighborhood $\mathcal{N}^*$ of Θ^* and $\tau_\epsilon > 0$ such that the iterates $\Theta^{(t)}$ from the MCEM algorithm initialized within $\mathcal{N}^*$ satisfies:*

$$P(||\Theta^{(t)} - \Theta^*||_2 < \epsilon \text{ for some } t \leq \tau_\epsilon) \to 1 \text{ as } \underline{m}, \overline{m}, b \to \infty.$$

5 Results

We consider several choice models for predicting customer responses starting from the naivest empirical cdf based multinomial model to multinomial probit models using different customer similarity networks. All the models were estimated on the calibration dataset containing 1896 observations randomly selected

from the dataset described in Section 2. Each model's parameter estimates were then used to calculate the likelihood and the Mean Average Deviation (MAD) in the validation dataset that contains 300 observations. In all cases, the covariance matrix of the utility error terms included separate variances for each vehicle model and off-diagonal elements that were set to zero. Table 2 reports the Mean Average Deviation (MAD) (for calibration and validation samples) and the log-likelihood in the holdout sample. We report performance of (a) the Naive Empirical approach that uses marginal purchase probabilities (without price and other covariate information) from the training dataset to predict consumer responses, (b) a simple multinomial probit model with all the covariates, (c) the Random coefficient probit model with random intercepts and slope coefficients for each zip-code, as well as the four different MNP models that shrinks based on the following network structures on the intercepts (preference) and slope (response) coefficients respectively (d) $(W = \mathrm{W}^{\mathrm{G}}, T = \mathrm{W}^{\mathrm{G}})$ (e) $(W = \mathrm{W}^{\mathrm{V}}, T = \mathrm{W}^{\mathrm{V}})$ (f) $(W = \mathrm{W}^{\mathrm{V}}, T = \mathrm{W}^{\mathrm{G}})$ (g) $(W = \mathrm{W}^{\mathrm{G}}, T = \mathrm{W}^{\mathrm{V}})$, where W and T are the weight matrices in the MWMNP model in (4)-(5) and W^{V} and W^{G} are defined in Section 2.1.

Parameters for the RCP are zipcode-specific and distributed multivariate normal over the population. The parameter vector can be conveniently divided into two components, each corresponding to the intercepts and the marketing mix parameters, respectively. Each of the models in (d)–(g) is based upon which of the two contiguity matrices W^{V} and W^{G} influences which component. This allocation guides our labeling strategy so that, the model in (f) for example, is labeled as $(\mathrm{W}^{\mathrm{G}}, \mathrm{W}^{\mathrm{V}})$, to indicate that the network weights in `MWMNP` for the intercepts is based on vehicle similarity and those for the marketing mix parameters on the geographic closeness between consumers. Note, that while $(\mathrm{W}^{\mathrm{G}}, \mathrm{W}^{\mathrm{G}})$ and $(\mathrm{W}^{\mathrm{V}}, \mathrm{W}^{\mathrm{V}})$ MNP models use a single customer similarity structure, they can be well estimated by the `MWMNP` algorithm.

Table 2: Training and test performances of different choice models

Model		**Calibration sample**	**Validation sample**		
		MAD	MAD	Log likelihood	# Parameters
Empirical		0.786	0.791		
Simple MNP		0.653			
Random coefficient probit		0.617	0.628	-387.76	16
MNP with different Network Weights	$(\mathrm{W}^{\mathrm{G}}, \mathrm{W}^{\mathrm{G}})$	0.641	0.597	-380.49	16
	$(\mathrm{W}^{\mathrm{V}}, \mathrm{W}^{\mathrm{V}})$	0.635	0.601	-364.06	16
	$(\mathrm{W}^{\mathrm{V}}, \mathrm{W}^{\mathrm{G}})$	0.633	0.598	-361.73	16
	$(\mathrm{W}^{\mathrm{G}}, \mathrm{W}^{\mathrm{V}})$	0.640	0.599	-373.07	16

In both samples, the homogeneous simple MNP and the RCP fit the data significantly better than the naive model. All MWMNP models fit the validation sample, but not the calibration sample, much better than the RCP, which suggests that the RCP is prone to over-fitting. We also estimated a version of the RCP that included the fuel efficiency of the previously owned vehicle as an additional variable. This model yielded a lower MAD (0.611) than the RCP in the calibration sample, though validation sample results were mixed, slightly superior MAD (0.626) but inferior log-likelihood (-391.69). The rankings of the MWMNP are the same for all three different fit metrics, from best to worst: 1. $(\mathrm{W}^{\mathrm{V}}, \mathrm{W}^{\mathrm{G}})$, 2. $(\mathrm{W}^{\mathrm{V}}, \mathrm{W}^{\mathrm{V}})$, 3. $(\mathrm{W}^{\mathrm{G}}, \mathrm{W}^{\mathrm{V}})$, and 4. $(\mathrm{W}^{\mathrm{G}}, \mathrm{W}^{\mathrm{G}})$. The positive difference in the log-likelihood of the validation sample between the $(\mathrm{W}^{\mathrm{V}}, \mathrm{W}^{\mathrm{V}})$ and $(\mathrm{W}^{\mathrm{G}}, \mathrm{W}^{\mathrm{G}})$ MNP models demonstrate that similar previous vehicle mpg captures more information on similarity among customers than their geographic proximity.

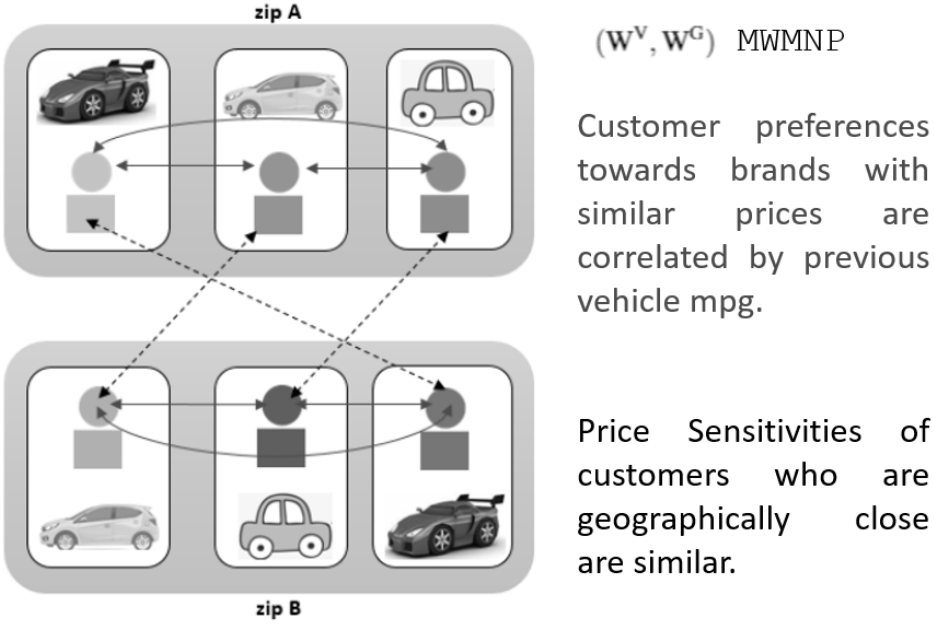

Figure 6: Schematic for the best fitting $(\mathrm{W}^{\mathrm{V}}, \mathrm{W}^{\mathrm{G}})$ MWMNP model.

Next, we concentrate on the best performing MWMNP model that uses $(\mathrm{W}^{\mathrm{V}}, \mathrm{W}^{\mathrm{G}})$ as the network weights for shrinking the intercepts and the slope coefficients respectively. Figure 6 shows the schematic for this MWMNP model which suggests that customer's price sensitivity is related through their geographic proximity but their preferences towards different brands (and in particular hybrid alternatives) with similar price is related by the fuel efficiency of their previous vehicle. Table 3 reports the coefficients for

the best fitting MWMNP model. Note that, the coefficient for the *Net Price* variable (price less rebate less Promotional incentive) is negative. This shows that the model though quite complex captures the correct sign of price sensitivity. The coefficients for the endogeneity correction terms (residual and price residual interaction) are statistically insignificant (significant), perhaps because (even though) the choice consists of a set of relatively homogeneous products from the same category. Strong firm-specific inertia effects are evidenced by the large and statistically significant coefficient for the *Last Make* variable, which shows that consumers tend to re-purchase cars made by the same manufacturer (Toyota, Honda).

Table 3: Summary of the coefficients from the fitted (W^V, W^G) MWMNP model

		Average Preference	Sigma
Intercepts/ Preference	Civic	0.465	1.000
	Prius	0.475	0.864
	Corolla	0.241	0.982
	Sentra	0.083	1.084
	Civic Hybrid	-0.481	1.522
	Scion xB	-0.335	1.239
	BMW m3	0.068	0.799

		Coefficients	Standard Error
	Net Price (in 100 thousand $)	-5.054	
Slopes/	Residual	0.280	
Responses	Net Price*residual	-1.558	
	Last make	7.681	

To provide some insights into the how the MWMNP model incorporates smoothing, we present two sets of analysis. The first pertains to the preference portion of the parameter vector (spatially correlated through W^V), specifically focusing on the preference coefficient for the Prius relative to the Corolla and Sentra respectively. Figure 7 plots these relative preferences from the (W^V, W^G) MWMNP model as previous vehicle mpg varies. Both graphs in Figure 7 clearly reveal that the higher the fuel efficiency of the consumer's currently owned vehicle the greater their preference for the Prius. Moreover, they also illustrate the non-linear relationship between previous vehicle mpg and preference for the hybrid alternative in Toyota Prius.

Figure 8 presents the next analysis, which shows a heat map of the price coefficient (spatially correlated through W^G) corresponding to the customers living within the area marked by the blue hashed lines

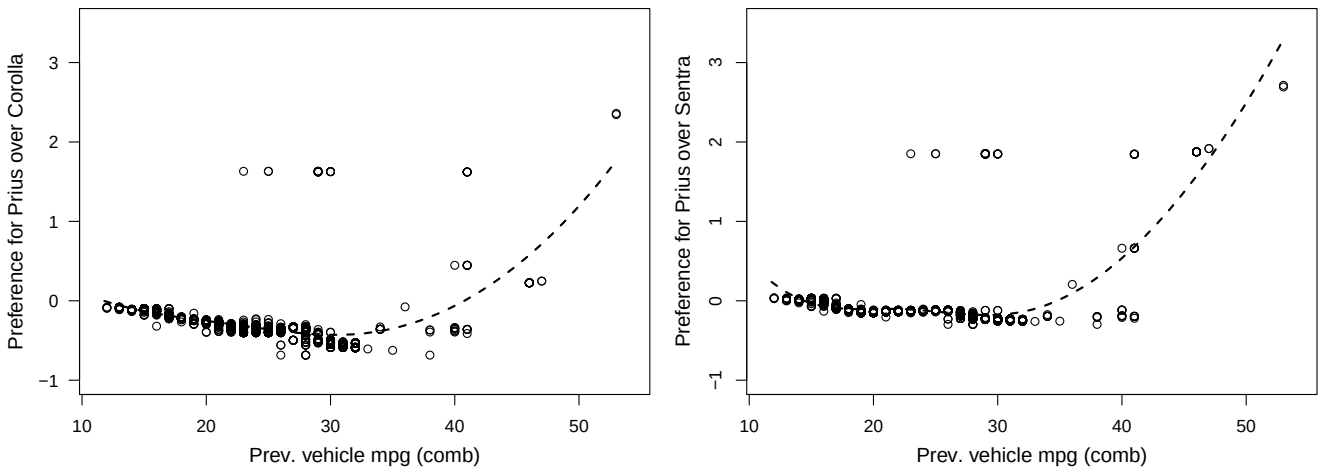

Figure 7: Estimated preference for Toyota Prius over Toyota Corolla and Nissan Sentra for different previous vehicle MPG values in the (W^V, W^G) model.

in Figure 2. Geographical zones corresponding to areas of low and high price sensitivity are evident and clearly reveal the local smoothing aspect of the `MWMNP`model.

We should discuss the preference and response coefficients here.

6 Managerial implications: Market Structure & Optimal Rebating

The fit statistics provide some insights into the relative performance of different models and the analysis presented in Figures 7 and 8 give a sense of how the `MWMNP` works. However, a complete picture of the model's advantage is best seen by examining its effect on a managerially relevant performance outcomes.

One important aspect of a model are the diagnostics that it can provide. In the current context we study its portrayal of the prevailing competitive market structure as revealed by the the Competitive Clout and Vulnerability measures that the model provides. The first step in this process is to derive each elements of the complete seven-by-seven cross-price elasticity matrix, whose elements represents the effect of a 1% price change of the column vehicle on the percentage change in share of the row vehicle. Table 4 shows the price elasticities from the RCP and the proposed (W^V, W^G) `MWMNP` model. In other words, the diagonal elements of the matrix are the "own-price" elasticities and the off-diagonals are the "cross-price"

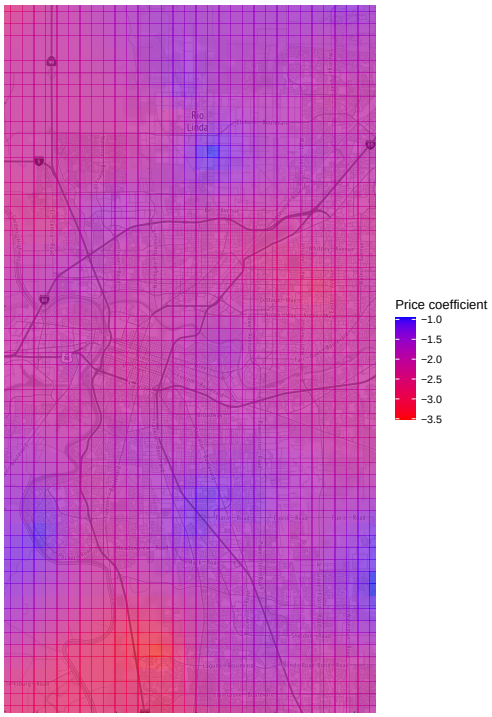

Figure 8: Estimated response coefficient for price in the (W^V, W^G) `MWMNP`model in the greater Sacramento area.

Table 4: Price elasticities of the different brands based on the RCP and multi-weight MNP models.

Price elasticities based on (W^V, W^G) multi-weight MNP model							
	Civic	Prius	Corolla	Sentra	Civic Hybrid	Scion xB	BMW m3
Civic	0.52	-0.18	-0.31	-0.33	-0.74	-0.22	-0.43
Prius	-0.15	1.11	-0.57	-0.29	-0.18	-0.48	-0.56
Corolla	-0.12	-0.28	1.08	-0.24	-0.12	-0.44	-0.42
Sentra	-0.07	-0.06	-0.15	0.98	-0.06	-0.07	-0.22
Civic Hybrid	-0.08	0.03	-0.03	-0.03	0.96	0.05	-0.04
Scion xB	-0.03	-0.05	-0.15	-0.07	-0.03	1.71	-0.11
BMW m3	-0.06	-0.06	-0.15	-0.13	-0.07	-0.06	1.70

Price elasticities based on RCP model							
	Civic	Prius	Corolla	Sentra	Civic Hybrid	Scion xB	BMW m3
Civic	0.88	-0.08	-0.31	-0.36	-0.46	-0.28	-0.75
Prius	-0.12	1.06	-0.65	-0.12	-0.05	-1.12	-0.46
Corolla	-0.13	-0.24	1.35	-0.18	-0.07	-0.78	-0.41
Sentra	-0.14	-0.05	-0.17	0.75	-0.07	-0.21	0.05
Civic Hybrid	-0.31	0.00	-0.12	-0.12	0.89	-0.13	-0.25
Scion xB	-0.04	-0.15	-0.26	-0.07	-0.03	2.28	-0.04
BMW m3	-0.22	-0.15	-0.31	0.16	-0.11	-0.10	1.70

elasticities. Define, *Competitive Clout of Brand A* as

$$\sum_{j \neq A} \text{percent decrease in the market share of Brand j if Brand A decreases price by } 1\%,$$

and the *Vulnerability of Brand A* as

$$\sum_{j \neq A} \text{percent decrease in the market share of Brand A if Brand j decreases price by } 1\%.$$

To compute the clout and vulnerability from Table 4, the off-diagonal elements are summed column- and row-wise, respectively, to yield each vehicle's Competitive Clout (its effect on others) and the its Vulnerability (others effect on it).

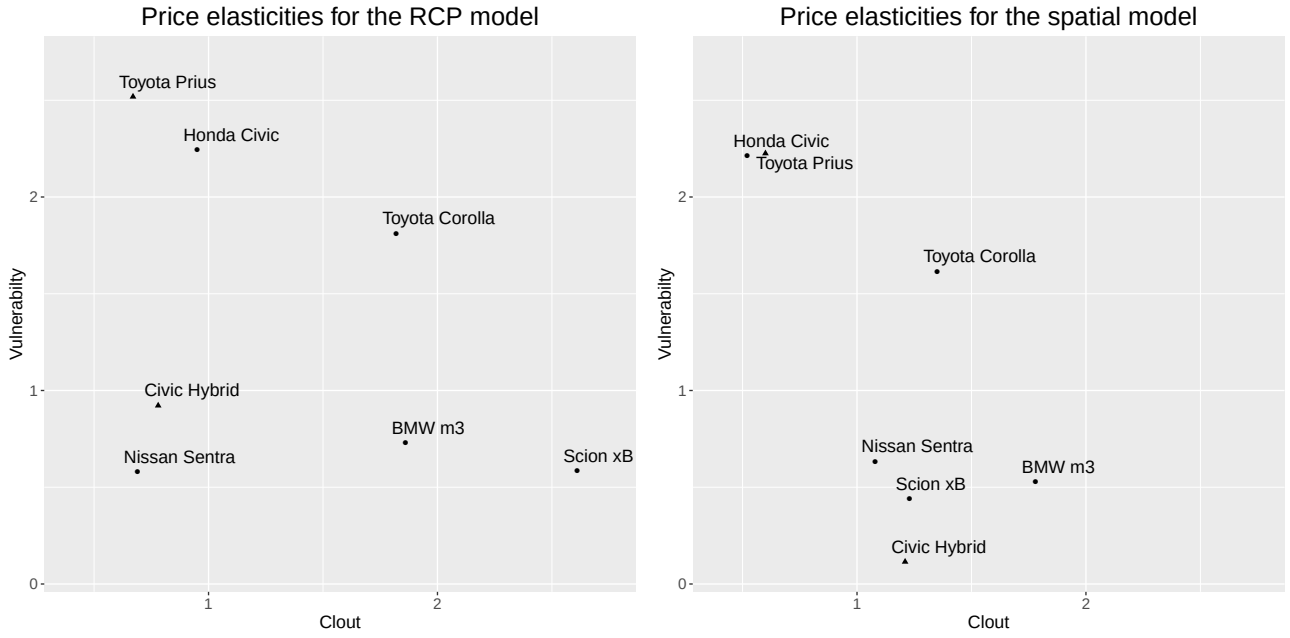

Figure 9: Estimated competitive clout and the vulnerability of the vehicle models using the RCP model and the $(\mathrm{W}^{\mathrm{V}}, \mathrm{W}^{\mathrm{G}})$ `MWMNP` model.

Figure 9 shows the Clout-Vulnerability maps from the RCP and best-fitting `MWMNP` model, which reveal some key differences in how the two maps depict the competitive market structure. The RCP model assigns a higher vulnerability to the Toyota Prius relative to the $(\mathrm{W}^{\mathrm{V}}, \mathrm{W}^{\mathrm{G}})$ `MWMNP` model. Also the other hybrid car model, the Honda Civic, has a much higher clout and lower vulnerability in the $(\mathrm{W}^{\mathrm{V}}, \mathrm{W}^{\mathrm{G}})$

map than the RCP map. The four car models Nissan Sentra, Scion XB, BMW m3 and Civic Hybrid are relatively similar in clout and vulnerability per the spatial model though the RCP model positions them much further apart in the Clout Vulnerability space. Finally, the BMW m3, despite having a small market share does not show a large difference in its competitiveness or vulnerability between the two models, perhaps because of BMW's brand recognition.

From the viewpoint of the Toyota Prius manager, the competitive market structure is crucially important when resources have to be allocated in a way that targets the early adopters most effectively. Since managers often operate with tight promotion budgets, small differences in model performance and competitive inference can translate into very meaningful difference in performance outcomes such as profits. To introspect whether differences in the market structure between the RCP and the developed `MWMNP` models as exhibited in Figure 9 and described before, we apply both models to the practical problem of designing a targeted promotion program to stimulate demand among such consumers.

The incentive program used to illustrate these differences is called Conquest Cash, which is a popular promotional tool used by US automakers to target owners and lessees of competing vehicles with a view to challenge brand loyalty and increase market share by "conquesting" owners of other manufacturers (as opposed to retaining current buyers) [1]. We take the position of the Prius marketing manager tasked with using a total promotional budget of P_B to target 100 out of the 300 customers in the validation sample in a way that maximizes profits. We consider $P_B = \$16000$ as the median rebate in the data was $ 160. To make the targeting "fair" each customer is constrained to get the same face value for the conquest cash incentive, i.e. $160. Each model, in turn, is used for the allocation task and the final comparison between the models is based on the net profits realized.

The basic building block of this analysis is the expected contribution EC_Z than can be realized from a consumer l in ZIP code Z when offered a rebate of $ 160 on the Toyota Prius. This contribution can be calculated as

$$EC_Z(l) = \text{Prob}(\text{customer } l \text{ buys a Prius}|\ \text{Prius Rebate is } \$160) \times \text{Margin}\big(\text{Net Price of Prius} - \$160\big).$$

[1] https://www.carsdirect.com/deals-articles/what-are-conquest-cash-incentives

Because data on manufacturer margin is not readily available, based on our discussions with industry experts we assume it to be 25% of the selling price. The total profit is obtained by summing these contributions over all of the consumers in the validation sample such that the 100 targeted customers get a rebate of $160 while the remainder get $0.

Table 5: Profit from conquest cash rebates based on different choice models

Model	**Profit (in $)**	**Incremental profit relative to RCP ($)**
RCP	383265.4	
(W^G, W^G)	384402.5	1137.1
(W^V, W^V)	396044.6	12779.2
(W^V, W^G)	409408.3	26142.9
(W^G, W^V)	379863.3	-3402.1

Table 5 reports the net profit obtained from the entire validation sample using each model, in turn, to determine the optimal set of targeted consumers. Profit based model performance matches the fit statistics in the holdout sample. Profits from the best-fitting MWMNP model (W^V, W^G) are about $26,000 more than the RCP and about $13,000 more than the next best MWMNP model (W^V, W^V). Overall this shows that the models lead to meaningfully different returns on investment.

To get a better sense of why these differences emerge we created a two-by-two matrix (see Table 6) showing the overlapping and non-overlapping customers targeted by the two models. Only 35 of the 100 targeted customers and 135 of the 200 non-targeted customers are common to both the RCP and (W^V, W^G) MWMNP models. The analysis shows that the vast majority of the consumers targeted by the two models (65%) are different and supports the idea that the local smoothing provided by the (W^V, W^G) MWMNP model helps it to target the right consumers and yield higher profits.

Table 6: Overlap between customers selected for rebates by the RCP and (W^V, W^G) MWMNP

		RCP	
		Rebate	No Rebate
(W^V, W^G)	Rebate	35	65
	No Rebate	65	135

We probe a little more deeply into these differences by examining the distribution of the mpg of of the vehicles previously owned by the 91 consumers in the three-digit Zipcode 956. The box-plot for is

distribution is shown in Figure 10 for (a) the entire group (b) the 19 consumers targeted for rebates by the RCP model and (c) the 20 consumers targeted for rebates by the proposed (W^V, W^G) MWMNP model. While the RCP targets people on the lower end of the mpg distribution, the (W^V, W^G) achieves its higher profits by focusing on consumers at the higher end of the mpg range.

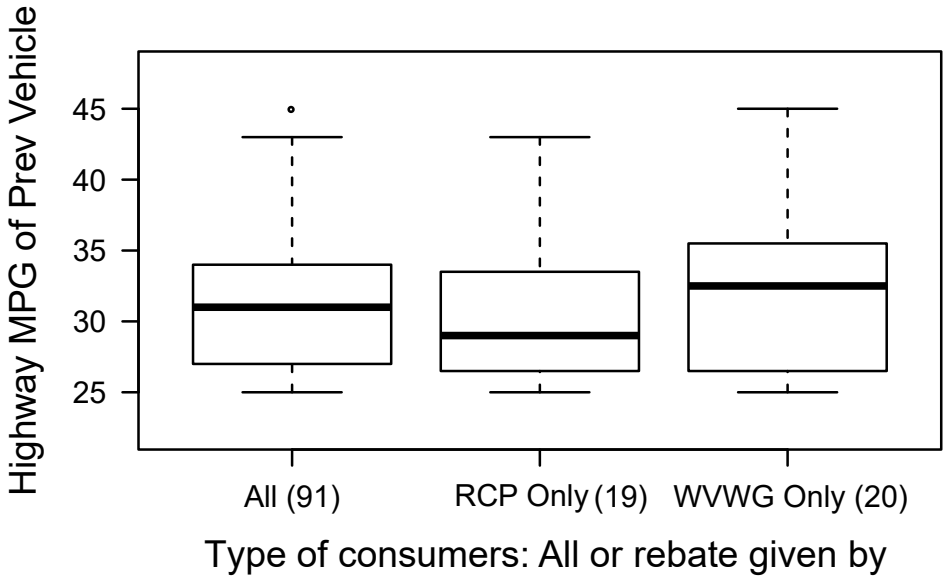

Figure 10: Boxplots for previous vehicle's highway mpg in the three digit ZIP code 956, for all consumers, consumers who only received under the RCP model, and consumers who only received under the (W^V, W^G) model.

7 Discussion

We develop and estimate a new discrete choice model that permits preference and response parameters to be correlated based on different network effects. We apply the model to transaction data from the Sacramento automobile market and find that the proposed model fits the data better than the RCP model and several benchmark models. While previous applications of spatial models in marketing have highlighted the role of geographic contiguity in demand prediction, we show that the similarity in consumers' previously purchased vehicles is more effective for estimating their preferences towards various brands of similar prices.

We make a methodological contribution to the weighted regression literature by proposing a new ap-

proach to simultaneously assimilate different kinds of contiguity matrices in multinomial probit model, which outperforms the existing approaches. We also expand on the existing promotional targeting literature by studying discounts based on previous purchases. We show how managers can use the model to come up with improved rebate programs that can accelerate adoption of hybrid cars. We studied data from the Sacramento market when hybrid cars had less than 1% of the new car market share and sales of new cars in the US was falling, which makes it extremely important for marketing managers of hybrid cars to efficiently target customers to make good use of scarce marketing dollars so that a new technological innovation is not side-lined.

In future, it will be interesting to study how the phenomena seen in the data evolve over time, i.e., how the customers preferences and price sensitivities evolve over time. Methodologically, it would involve incorporating dynamic shrinkage effects. Also, in our work contiguity is only based on geographic distance and previous vehicle information, which could be expanded to include other socio-demographic or purchase characteristics. For example, consumer's income, ethnicity, age or previous search history or dealership visits may form a viable basis of network matrices that explain variation in intrinsic preferences. Similarities from the previous vehicle-based contiguity matrix can be expanded to include more attributes, for example, engine size, gas mileage and so on. Statistically, this would require development of new methods that in presence of a large number of potentially useful network weight matrices can chose optimal network weights for different covariates. Finally, the modeling approach could be extended to accommodate spatial correlation in the features of the alternatives themselves, similar to the spatial demand model of Duan and Mela (2009) in which geographic distance between alternatives (outlet location) serves to identify consumer preferences.

8 Supplementary Materials

8.1 Proof of Lemma 1

Consider the singular value decomposition of X:

$$X_{NC\times p} = U_{NC\times p} D_{p\times p} V_{p\times p}$$

where D is a diagonal matrix and V is a orthogonal matrix. Both D and V have full rank. Based on (12), given U and $r_x^{(l)}$, we can evaluate $r(t,l)$ in $\mathcal{O}(NC)$ for each t and l. Note that, computing $V'D^{-1}J(t,i)$ in (11) needs $O(p^3)$ and can be done only once for each t as it is invariant across l. Multiplying $V'D^{-1}J(t,i)$ and $r(t,l)$ has complexity $\mathcal{O}(p^2)$. Thus, computing $\{\check{\beta}_i^{(l)} : l = 1,\ldots,b\}$ knowing the SVD of X has complexity $\mathcal{O}(p^3 + b(NC+p^2))$ at the $(t+1)^{\text{th}}$ outer loop iteration. Similarly, computing $\{\check{\boldsymbol{\alpha}}_i^{(l)} : l = 1,\ldots,b\}$ is $\mathcal{O}(C^3 + b(NC+C^2))$.

For the E-step for each of the i machines: computing the mean and the variance is $\mathcal{O}(pC)$. But, this can be done once for any fixed t and i. Generating the random variables for the entire inner loop on the E-step is $\mathcal{O}(mC)$. Thus, the E-step computation for t^{th} outer loop is $\mathcal{O}(mC+pC)$ for each of the i machines.

At the t^{th} iteration of the M-step, for each of the i machines the complexity is $\mathcal{O}(\kappa^3C^3 + b(NC+\bar{c}^2))$ where, $\kappa = \max(1, p/C)$. Thus, the total complexity for t^{th} iterative step of the MCEM algorithm using n parallel machines is $\mathcal{O}(mC+pC+\kappa^3C^3+b(NC+\kappa^2C^2))$. The SVD of X is of complexity $\mathcal{O}(NCp^2)$ which needs to be done only once. So, the complexity for running τ iterations of the outer loop is

$$\mathcal{O}\left(NC\left(p^2 + \tau\left[b + \frac{m}{N} + \kappa^3\frac{C^2}{N} + b\kappa^2\frac{C}{N}\right]\right)\right).$$

8.2 Computation time and performance

Bikram: Can you report the τ (convergence time) in each cases; the set-up of the parameters; the mean square errors; In table 7, we report the average computational time as N and C varies and $m = 300$, $\underline{m} = 50$, $\overline{m}$ 30 (after thinning) and $b = 100$. How fast does the outer loop converge? – WILL DO.

Machine is intel core I7 3.4 Ghz, only 16 GB of RAM.

The R code that reproduces our simulation results can be downloaded from the following link: https://github.com/gmukherjee.

Table 7: Time in minutes averaged over 100 simulations as sample size N and C varies

C	N		
	1000	**2000**	**5000**
3	11.36	31.93	51.52
5	19.81	39.36	99.93
10	31.07	62.53	168.25

8.3 Proof of Lemma 2

Define $\boldsymbol{y}_i^{(k)} = \{y_{ic}^{(k)} : c = 1, \ldots, C\}$. For any $\boldsymbol{x} \in \mathbb{R}^C$, consider the energy function $g(\boldsymbol{x}) = |\boldsymbol{x}|_1$. Note that, g is unbounded off compact sets. For any fixed i, note that the Markov chain $(\boldsymbol{y}_i^{(k)})_{k\geq 1}$ is Feller continuous. Consider the conditional expectation: $\mathbb{E}\{g(\boldsymbol{y}_i^{(k)})|\boldsymbol{y}_i^{(k-1)}\}$ based on which, we will establish the following drift condition on the energy function g

$$\mathbb{E}\{g(\boldsymbol{y}_i^{(k)})|\boldsymbol{y}_i^{(k-1)}\} \leq \rho_i \, g(\boldsymbol{y}_i^{(k-1)}) + L_i \, , \tag{14}$$

where, ρ_i and L_i are universal constant independent of k and $\rho_i < 1$ for all i. Once the above drift condition is proved, it follows from Theorem 15.2.8 of Meyn and Tweedie (2012) that the Markov chain in geometrically ergodic. Next, we prove (14). For that purpose, with out loss of generality we assume the following:

- $\alpha_{ic}^{(t)} + \boldsymbol{x}_{ic}'\beta_i^{(t)}$ is 0 and $\sigma_c = 1$ for all $c = 1, \ldots, C$,
- the first alternative was chosen by customer i, i.e., $O_i = 1$.

Define the threshold $t_i^{(k)} = \max_{c\neq 1} y_{ic}^{(k)}$. The following properties are obeyed in our iterative algorithm:

(a) $y_{i1}^{(k+1)}$ is generated from a standard normal random variable that is truncated between $[t_i^{(k)}, \infty)$,

(b) $\{y_{ic}^{(k+1)} : c = 2, \ldots, C\}$ are generated from standard normal random variables that are truncated between $(-\infty, y_{i1}^{(k+1)}]$.

(c) conditioned on $y_{i1}^{(k+1)}$, the utilities in $\{y_{ic}^{(k+1)} : c = 2, \ldots, C\}$ are independent among themselves,

(d) The dependence between $y_{i1}^{(k+2)}$ and $\{y_{ic}^{(k+1)} : c = 1, \ldots, C\}$ is only through $t_i^{(k+1)}$.

Now, note that by property (b) it follows that $g(\boldsymbol{y}_i^{(k)}) \leq C\,[y_{i1}^{(k)}\, 1\{y_{i1}^{(k)} > 0\} + |\min_{c\neq 1} y_{ic}^{(k)}|]$. Thus,

$$C^{-1}\,\mathbb{E}\{g(\boldsymbol{y}_i^{(k)})|\boldsymbol{y}_i^{(k-1)}\} \leq \mathbb{E}\big[y_{i1}^{(k)} 1\{y_{i1}^{(k)} > 0\}\big|\boldsymbol{y}_i^{(k-1)}\big] + \mathbb{E}\big[\,|\min_{c\neq 1} y_{ic}^{(k)}|\,\big|\boldsymbol{y}_i^{(k-1)}\big]\,. \tag{15}$$

We first concentrate on the first term in the right side of (15). By property (b) and (c), it follows that for each $k \geq 1$, the distribution of $t_i^{(k)}$ is stochastically dominated by Z^* which is the maximum of $C-1$ independent standard normal random variables. Note that for $k = 0$, $t_i^{(0)} \stackrel{d}{=} Z^*$. Thus, the first term in the right side of (15) is bounded above by $\mathbb{E}_{Z^*}T_+$ where, T_+ is the positive part of the truncated normal random variable, truncated on $[Z^*, \infty)$. This can be further simplified as:

$$E_1 = \mathbb{E}_{Z^*}T_+ = \int \phi(u)(1-\Phi(u))^{-1} f(u)\,du$$

where, f is the density of Z^*; ϕ and Φ are standard normal pdf and cdf respectively.

Next, we consider the second term on the right side of (15). By property (b) and (c), the conditional distribution of $[|\min_{c\neq 1} y_{ic}^{(k)}|\,|\boldsymbol{y}_i^{(k-1)}]$ is stochastically dominated by the distribution of

$$\check{Z}^* + \big[y_{i1}^{(k)} 1\{y_{i1}^{(k)} > 0\}\big|\boldsymbol{y}_i^{(k-1)}\big]\,,$$

where $\check{Z}^*$ is the maximum of $C-1$ i.i.d. truncated normal random variables with mean 0, standard deviation 1 and truncated on $[0, \infty)$. Thus, it follows that

$$\mathbb{E}\{g(\boldsymbol{y}_i^{(k)})|\boldsymbol{y}_i^{(k-1)}\} \leq C(2E_1 + \mathbb{E}\check{Z}^*)\,.$$

The terms on right side above is always bounded above as they are independent of k and only depend on

C. Thus, we have established the drift condition in (14) which completes the proof of the lemma.

8.4 Proof of Theorem 1

We first show that under assumption 1, for large b, conditioned on $\boldsymbol{\mu}_i^{(t+1)}$ the backfitting iterates of the preference and response parameters converges in the M-step for each iteration t of the outer loop of the proposed MCEM algorithm. For this purpose, we next show that assumption 1 implies that $||\mathcal{P}^{(t)}\mathcal{S}^{(t)}||_\infty < 1$ for all t.

By $\mathbb{W}_i^{(t)}$ and $\mathbb{T}_i^{(t)}$ denote the $NC \times NC$ diagonal matrices used in the tth iterative step of the MCEM algorithm:

$$\mathbb{W}_i^{(t)} = (W_i \otimes I_C)(I_n \otimes (\Sigma^{(t)})^{-1})(W_i \otimes I_C) = (W_i^2 \otimes I_C)(I_n \otimes (\Sigma^{(t)})^{-1}) = W_i^2 \otimes (\Sigma^{(t)})^{-1} \tag{16}$$

$$\mathbb{T}_i^{(t)} = (T_i \otimes I_C)(I_n \otimes (\Sigma^{(t)})^{-1})(T_i \otimes I_C) = (T_i^2 \otimes I_C)(I_n \otimes (\Sigma^{(t)})^{-1}) = T_i^2 \otimes (\Sigma^{(t)})^{-1} \, . \tag{17}$$

Note that, $\mathbb{P}_i^{(t)} = (Z'\mathbb{W}_i^{(t)}Z)^{-1}Z'\mathbb{W}_i^{(t)}$. Again, as the L_2 norm of each row of W equals 1, we have, $(Z'\mathbb{W}_i^{(t)}Z)^{-1} = \Sigma^{(t)}$ and so,

$$\mathbb{P}_i^{(t)} = \Sigma^{(t)} * (\boldsymbol{w}_i^2 \otimes (\Sigma^{(t)})^{-1}) = \boldsymbol{w}_i^2 \otimes I_C \, ,$$

where, $\boldsymbol{w}^2$ is the $1 \times N$ row vector representation of the diagonal matrix W_i^2.

Recall, $\mathcal{P}_i^{(t)} = (I_C - C^{-1}\mathbf{1}\mathbf{1}')\mathbb{P}_i^{(t)}$ and $\mathcal{S}_i^{(t)} = (I_C - C^{-1}\mathbf{1}\mathbf{1}')X_i\mathbb{S}_i^{(t)}$. Thus, $\mathcal{P}_i^{(t)} = \bar{\boldsymbol{w}}_i^2 \otimes I_C$, where $\bar{\boldsymbol{w}}_i$ is a $1 \times N$ row vector whose lth element is $W_{il}^2 - C^{-1}$. Define, the $N \times N$ matrix, $\bar{W}^{(2)} = W \circ W - C^{-1}\mathbf{1}_N\mathbf{1}_N'$. Note that, the $NC \times NC$ square matrices $\mathcal{P}^{(t)}$ and $\mathcal{S}^{(t)}$ are respectively derived by binding $\{\mathcal{P}_i^{(t)} : i = 1, \ldots, N\}$ and $\{\mathcal{S}_i^{(t)} : i = 1, \ldots, N\}$ row wise. Thus, we have for any $t \geq 1$,

$$||\mathcal{P}^{(t)}\mathcal{S}^{(t)}||_\infty = \sup_{i=1,\ldots,N} ||\bar{W}^{(2)}\mathcal{S}_i^{(t)}||_\infty \, . \tag{18}$$

Note that, $\mathcal{S}_i^{(t)} = J_i X(X'T_i^{(t)}X)^{-1}X'T_i^{(t)} := J_i T_i^{(2)}$ where J_i is the $C \times NC$ matrix: $J_i = e_i \otimes (I_C - C^{-1}\mathbf{1}\mathbf{1}')$ and e_i is the unit vector in $\mathbb{R}^N$ for the ith coordinate. If the condition number of $\Sigma^{(t)}$ is bounded

above by L then (18) is further simplified as

$$||\mathcal{P}^{(t)}\mathcal{S}^{(t)}||_\infty \leq L \sup_{i=1,\ldots,N} || \bar{W}^{(2)} J_i T_i^{(2)}||_\infty ,$$

with the bound on the right side now being independent of t. From (13) of Assumption 1, it follows that the right side above is less than 1 which subsequently implies that $||\mathcal{P}^{(t)}\mathcal{S}^{(t)}|| < 1$ for all t. Now, by corollary 4.3 of Buja et al. (1989), it follows that the backfitting equations used in the M-step are consistent for any $\boldsymbol{\mu}_i^{(t+1)}$. Thus, as $b \to \infty$, the iterates in the M-step converges to the maximum value of the likelihood given by the model in (3)-(5). The remainder of the proof regarding optimality of iterates produced in the outer loop of the MCEM algorithm follows directly from Theorem 5 of Neath (2013) which is a modified version of Theorem 1 of Chan and Ledolter (1995).

References

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

Anselin, L. (2013). *Spatial econometrics: methods and models*, Volume 4. Springer Science & Business Media.

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2014). *Hierarchical modeling and analysis for spatial data*. CRC press.

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, 841–890.

Bucklin, R. E., S. Siddarth, and J. M. Silva-Risso (2008). Distribution intensity and new car choice. *Journal of Marketing Research 45*(4), 473–486.

Buja, A., T. Hastie, and R. Tibshirani (1989). Linear smoothers and additive models. *The Annals of Statistics*, 453–510.

Carlier, M. (2021). Light vehicle retail sales in the united states from 1976 to 2020.

Carroll, R. J. and D. Ruppert (1988). *Transformation and weighting in regression*, Volume 30. CRC Press.

Chan, K. and J. Ledolter (1995). Monte carlo em estimation for time series models involving counts. *Journal of the American Statistical Association 90*(429), 242–252.

Chintagunta, P., J.-P. Dubé, and K. Y. Goh (2005). Beyond the endogeneity bias: The effect of unmeasured brand characteristics on household-level brand choice models. *Management Science 51*(5), 832–849.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 39*(1), 1–22.

Duan, J. A. and C. F. Mela (2009). The role of spatial demand on outlet location and pricing. *Journal of Marketing Research 46*(2), 260–278.

Fotheringham, A. S., C. Brunsdon, and M. Charlton (2003). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.

Fotheringham, A. S., W. Yang, and W. Kang (2017). Multiscale geographically weighted regression (mgwr). *Annals of the American Association of Geographers 107*(6), 1247–1265.

Härdle, W. K., M. Müller, S. Sperlich, and A. Werwatz (2004). *Nonparametric and semiparametric models*. Springer Science & Business Media.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized additive models*, Volume 43. CRC press.

Heutel, G. and E. Muehlegger (2015). Consumer learning and hybrid vehicle adoption. *Environmental and resource economics 62*(1), 125–161.

Jank, W. and P. Kannan (2005). Understanding geographical markets of online firms using spatial models of customer choice. *Marketing Science 24*(4), 623–634.

Karmakar, B., O. Kwon, G. Mukherjee, S. Siddarth, and J. M. Silva-Risso (2021). Does a consumer's previous purchase predict other consumers' choices? a bayesian probit model with spatial correlation in preference. *under review in QME; available at:* `bit.ly/spatialprobit`.

Keane, M. P. (1994). A computationally practical simulation estimator for panel data. *Econometrica: Journal of the Econometric Society*, 95–116.

LeSage, J. P. (2004). A family of geographically weighted regression models. In *Advances in spatial econometrics*, pp. 241–264. Springer.

Li, Z., A. S. Fotheringham, W. Li, and T. Oshan (2019). Fast geographically weighted regression (fastgwr): a scalable algorithm to investigate spatial process heterogeneity in millions of observations. *International Journal of Geographical Information Science 33*(1), 155–175.

Mammen, E., O. Linton, and J. Nielsen (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics 27*(5), 1443–1490.

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

McCulloch, R. E., N. G. Polson, and P. E. Rossi (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.

McLachlan, G. J. and T. Krishnan (2007). *The EM algorithm and extensions*, Volume 382. John Wiley & Sons.

Meyn, S. P. and R. L. Tweedie (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Moore, A. G. (1991). *Crossing the chasm: marketing and selling technology products to mainstream customers*. HarperBusiness.

Moore, G. (2002). *Crossing the Chasm: Marketing and Selling Disruptive Products to Mainstream Customers*. Collins Business Essentials. HarperCollins.

Natarajan, R., C. E. McCulloch, and N. M. Kiefer (2000). A monte carlo em method for estimating multinomial probit models. *Computational statistics & data analysis 34*(1), 33–50.

Neath, R. C. (2013). On convergence properties of the monte carlo em algorithm. In *Advances in modern statistical theory and applications: a Festschrift in Honor of Morris L. Eaton*, pp. 43–62. Institute of Mathematical Statistics.

Opsomer, J. D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis 73*(2), 166–179.

Opsomer, J. D. and D. Ruppert (1997). Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics 25*(1), 186–211.

Petrin, A. and K. Train (2010). A control function approach to endogeneity in consumer choice models. *Journal of marketing research 47*(1), 3–13.

Robertson, T. S. (1967). The process of innovation and the diffusion of innovation. *Journal of marketing 31*(1), 14–19.

Rogers, E. M. (2010). *Diffusion of innovations*. Simon and Schuster.

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

Rossi, P. E., G. M. Allenby, and R. McCulloch (2012). *Bayesian statistics and marketing*. John Wiley & Sons.

Roy, V. and J. P. Hobert (2007). Convergence rates and asymptotic standard errors for markov chain monte carlo algorithms for bayesian probit regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(4), 607–623.

Ruppert, D. and M. P. Wand (1994). Multivariate locally weighted least squares regression. *The annals of statistics*, 1346–1370.

Tan, Z. and C.-H. Zhang (2019). Doubly penalized estimation in additive regression with high-dimensional data. *The Annals of Statistics 47*(5), 2567–2600.

Tosetti, E. and V. Vinciotti (2019). A computationally efficient correlated mixed probit for credit risk modelling. *Journal of Royal Statistical Society, Series A*, 1183–1204.

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

Wei, G. C. and M. A. Tanner (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association 85*(411), 699–704.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *The Journal of Human Resources 50*(2), 420–445.

Yang, S. and G. M. Allenby (2003). Modeling interdependent consumer preferences. *Journal of Marketing Research 40*(3), 282–294.

Zettelmeyer, F., F. S. Morton, and J. Silva-Risso (2006). How the internet lowers prices: Evidence from matched survey and automobile transaction data. *Journal of marketing research 43*(2), 168–181.