WORKING PAPERS

N° 16-741

December 2016

"Models of Consumer Demand for Differentiated Products"

Timothy J. Richards and Celine Bonnet

Toulouse School of Economics

Models of Consumer Demand for Differentiated Products

Timothy J. Richards and Celine Bonnet*

September 12, 2016

Abstract

Advances in available data, econometric methods, and computing power have created a revolution in demand modeling over the past two decades. Highly granular data on household choices means that we can model very specific decisions regarding purchase choices for differentiated products at the retail level. In this chapter, we review the recent methods in modeling consumer demand, and their application to problems in industrial organization and strategic marketing.

Keywords: consumer demand, discrete choice, discrete-continuous choice, shopping basket models, machine learning.

JEL Codes: D43, L13, L81, M31.

*Richards is Professor and Morrison Chair of Agribusiness, Morrison School of Agribusiness, W. P. Carey School of Business, Arizona State University; Bonnet is Senior Researcher, Toulouse School of Economics and INRA, Toulouse, France. Contact author, Richards: Ph. 480-727-1488, email: trichards@asu.edu.

1 Introduction

Advances in available data, econometric methods, and computing power have created a revolution in demand modeling over the past two decades. Highly granular data on household choices means that we can model very specific decisions regarding purchase choices for differentiated products at the retail level. In this chapter, we review the recent methods in modeling consumer demand that have proven useful for problems in industrial organization and strategic marketing.

Analyzing problems in the agricultural and food industries requires demand models that are able to address heterogeneity in consumer choice in differentiated-product markets. Discrete choice models, for example, are particularly adept at handling problems that concern potentially dozens of choices as they reduce the dimensionality of product space into the smaller space occupied by product attributes. Discrete choice models, however, suffer from the independence of irrelevant alternatives (IIA) problem, so improvements on the basic logit model – the nested logit, mixed logit, and Bayesian versions of each – have been developed that are more relevant for consumer demand analysis, and a wide range of applied problems. Yet, the fundamental assumption that consumers make discrete choices among products remained unsatisfying for a large class of problems.

Beyond the well-understood problems with the logit model, there are many settings in which choices are not exactly discrete. Families that purchase several brands of soda, for example, make multiple discrete choices, as do consumers who purchase Sugar Pops for their children, and granola for themselves. Consumers who purchase a certain cut of beef make a discrete choice among the several they face, but then make a continuous choice as to how much to purchase (Dubin and McFadden 1984). Often, our interest lies more in the structure of the continuous part than the discrete part. Consumers also reveal a demand for variety when their purchase cycle is a week, anticipating 3 meals per day for the next 7 days, when purchasing food. This demand for variety is often manifest in multiple discrete-continuous decisions, each with a continuous quantity (Bhat 2005, 2008). In this chapter,

we will describe the evolution of demand model to describe more types of purchases, and more accurately describe the purchases observed in "the real world."

Developments in the spatial econometrics literature opened up an entirely new way of thinking of demand models (Pinkse, Slade and Brett 2002; Slade 2004). When we think about the demand for differentiated products, our notion of differentiation is all about distance, whether in geographic, attribute, demographic, or even temporal space. The differences between products can be expressed in terms of each definition of space. Most importantly for applied problems, writing demand models in terms of the spatial distance between products can potentially reduce a high-dimension problem to one that is more simple, and empirically tractable. We will briefly review the spatial econometrics literature, and the "distance metric" approach to demand estimation.

Finally, we address the frontier of demand analysis. Researchers working in "big data" have realized the power of machine learning methods to understand data patterns in largely atheoretic, but incredibly powerful ways (Varian 2014; Belloni et al. 2014; Bajari et al. 2014). Once limited to only forecasting and prediction, machine learning models have become increasingly important in econometric inference, again driven by the availability of massive data sets, both in terms of their depth (number of observations) and breadth (number of predictors).

We complete the chapter by suggesting some useful applications for new consumer demand models, such as empirical industrial organization, behavioral inference, and determining causality in natural experiments.

2 Models of Discrete Choice

When products are highly differentiated, the fundamental assumption of representative consumer models, namely that consumers buy a small proportion of each item in the data set, falls apart. Rather, with access to data on a highly disaggregate set of products, say at the UPC-level among ready-to-eat cereals or yogurt, it is more accurate to describe the decision

process as choosing only one brand from potentially dozens on offer. Building on the conceptual framework for discrete choices of items from Luce (1959), researchers in economics (McFadden 1974) and marketing (Guadagni and Little 1983) began to build a family of demand models that could describe purchases as discrete choices among differentiated items. Based on the assumption that preferences are randomly distributed among individuals, discrete choice models grew to become a standard approach to demand analysis due to their tractability, and their ability to reduce high-dimension problems to relatively simple estimation routines. In this section, we will describe the general model, the mixed logit, and other specific cases.

2.1 Models of Demand

Variation in choice among consumers is driven by the assumption that tastes are randomly distributed over individuals. Consider a consumer h who faces a set of J alternatives. For each alternative $j \in J$, he obtains a certain level of utility U_{hj}. The consumer i chooses the alternative j that gives him the highest utility: $\forall k \neq j, U_{hj} > U_{hk}$. Some attributes of the alternative j and some characteristics of the consumer h are observed and some others are not. Therefore, the indirect utility of the consumer i for the alternative j can be decomposed into two components: $U_{hj} = V_{hj} + \epsilon_{ij}$ where V_{hj} is a function of observed characteristics and ϵ_{hj} is a random term that captures unobserved factors. Different models are derived from the specification of the distribution of this error term.

The general model, namely the mixed logit model or the random coefficient logit model, that approximates any random utility model representing discrete choices (McFadden and Train 2000), can be specified as follows:

$$U_{hj} = \alpha_h p_j + \sum_k \gamma_{hk} b_{kj} + \xi_j + \epsilon_{hj} \tag{1}$$

where p_j is the price of the alternative j, α_h is the marginal utility of income for the consumer h, b_{kj} are the k^{th} observed attribute of the alternative j and γ_{hk} is the parameter associated with each observed variable that captures a consumer's tastes, ξ_j represents the unobserved

time invariant characteristics of the alternative j and ϵ_{hj} is the error term, which is assumed to follow a Type I Extreme Value distribution. The coefficients α_h and $\gamma_h = (\gamma_{h1}, \gamma_{h2}, ..., \gamma_{hk})$ vary over consumers with density $f(\alpha_h, \gamma_h)$,which, in most applications, is specified as normal or log-normal. The analyst then estimates the mean vector and covariance matrix of the random distribution.[1] The distribution $f(\alpha_h, \gamma_h)$ can also depend on observed characteristics of consumers (Bhat 2000), in which case the random coefficients are then specified as

$$\begin{pmatrix} \alpha_h \\ \gamma_h \end{pmatrix} = \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} + \Pi D_h + \Sigma \nu_h, \tag{2}$$

where α and γ are the mean marginal utility of income and the mean taste for characteristics respectively, Π is a matrix of coefficients that measure the taste for consumer according to their observed characteristics D_h, Σ is a matrix of coefficients that represent the variance of each additional unobserved characteristic ν_h and possible correlations between them.

Consumers choose the alternative that maximizes their utility. The individual probability of choosing the alternative j for the consumer h is given by:

$$\begin{aligned} s_{hj} &= P\left[U_{hj} > U_{hi}, \forall i = 1, ..., J, i \neq j | b_j, p_j, \xi_j, D_h, \nu_h\right] \\ &= \frac{\exp(\alpha p_j + b_j \gamma_h + \xi_j + [p_j, b_j](\Pi D_h + \Sigma \nu_h))}{\sum_{l=1}^{J} \exp(\alpha p_l + b_l \gamma_h + \xi_l + [p_l, b_l](\Pi D_h + \Sigma \nu_h))} \end{aligned}$$

and the aggregated probability, that is the market share of the alternative j, is

$$s_j = \int s_{hj} f(\nu_h) d\nu_h,$$

assuming the alternatives cover the entire market of interest. Below, we discuss alternatives for introducing an "outside option" to expand the definition of market share. Estimating market shares, however, are typically only relevant when used to estimate price elasticities.

2.2 Demand Elasticities

Price elasticities for the mixed logit reflect very general and flexible patterns of substitution among products (McFadden and Train 2000), and take the following form for the demand

[1] Triangular, uniform Rayleigh or truncated normal distributions are also used in the literature (Revelt and Train 2000, Hensher and Greene 2003, Siikamaki 2001, Revelt 1999).

of the alternative j with respect to the alternative k :

$$\eta_{jk} = \frac{\partial s_j}{\partial p_k}\frac{p_k}{s_j} = \begin{cases} \frac{p_j}{s_j}\int \alpha_h s_{hj}(1 - s_{hj}) f(\nu_h) d\nu_h \text{ if } j = k \\ -\frac{p_k}{s_j}\int \alpha_h s_{hj} s_{hk} f(\nu_h) d\nu_h \text{ if } j \neq k \end{cases} \tag{3}$$

When the price of the alternative j varies, the probability of choosing the other alternatives varies according to their attributes and the ones of the alternative j. Introducing consumer characteristics and unobserved individual components takes into account heterogeneity of consumers' preferences which, in turn, creates the flexibility we desire. However, one drawback of this method is that it lacks a closed form, so simulation methods are required to estimate all parameters, and obtain price elasticities. Although the mixed logit is the most general form, McFadden and Train (2000) show that all other forms of the logit model are, in fact, special cases of the mixed logit.

2.3 Particular cases

Two common discrete choice models can be derived from the mixed logit by imposing restrictions on the random variables describing consumer preferences: the simple logit and the nested logit. Constraining the random variables that describe unobserved heterogeneity permit closed form expressions for choice probabilities, and for price elasticities.

The simple logit model differs from the mixed logit in that the parameters are assumed to be fixed: $\alpha_h = \alpha$ and $\gamma_h = \beta$. With this assumption, the aggregated choice probabilities are then written as the logit expression:

$$s_j = \frac{\exp(\alpha p_j + \sum_l \gamma_l b_{lj} + \xi_j)}{\sum_{l=1}^{J} \exp(\alpha p_l + \sum_l \gamma_l b_{lj} + \xi_k)} \tag{4}$$

and the price elasticities become more tractable:

$$\eta_{jl} = \begin{cases} \alpha p_j (1 - s_j) \text{ if } j = l \\ -\alpha p_l s_l \text{ if } j \neq l \end{cases} \tag{5}$$

The cross-price elasticities of the alternative j with respect to the alternative l only depends on the alternative l whatever the alternative j considered. Therefore, when the price of an alternative changes, the share of each other alternative is affected in exactly the same way. Moreover, this simple model exhibits the IIA property referred to above as the ratio of probabilities of two alternatives j and l is independent from changes in the price of other alternatives. Although restrictive, the IIA property provides a very convenient form for the choice probabilities, which also explains its popularity.

Some additional assumptions on the distribution of the error term generate another closed-form expression for choice probabilities, and offer more flexibility in substitution patterns than the simple logit. In particular, when the set of alternatives can be decomposed into several subsets, and alternatives within each subset are correlated in demand, the nested logit results. In the nested logit, the IIA property holds for alternatives belonging to the same group, but does not hold for alternatives in different subsets. Assuming each alternative belongs to a group $g \in \{1, ..., G\}$, the number of alternatives within each group g is J_g and the error term can be written as $\epsilon_{ij} = \zeta_{ig} + (1 - \sigma_g)\upsilon_{ij}$ where υ_{ij} follows a Type I Extreme Value distribution, ζ_{ig} is common to all alternatives of the group g and has a cumulative distribution function that depends on σ_g, with $\sigma_g \in [0, 1]$. Importantly, the parameter σ_g measures the degree of correlation between alternatives within the group g. When σ_g tends toward 1, preferences for alternatives of the group g are perfectly correlated, meaning that the alternatives are perceived as perfect substitutes. When σ_g tends toward 0 for all $g = 1, ...G$, the nested logit model is equivalent to the simple logit model.

In the nested logit model, the analytical expression for the choice probabilities is:

$$\begin{aligned} s_j &= s_{j/g} s_g \text{ where } s_{j/g} = \frac{\exp(\frac{\alpha p_j + \sum_l \gamma_l b_{lj} + \xi_j}{1-\sigma_g})}{\frac{\exp(I_g)}{1-\sigma_g}} \text{ and } s_g = \frac{\exp(I_g)}{\exp(I)} \\ \text{with } I_g &= (1 - \sigma_g) \ln \left(\sum_{j \in J_g} \exp(\frac{\alpha p_j + \sum_l \gamma_l b_{lj} + \xi_j}{1 - \sigma_g}) \right) \text{ and } I = \ln \sum_{g=1}^{G} \exp(I_g). \end{aligned} \tag{6}$$

The nested logit model is more general and substitution patterns are more flexible than the multinomial logit model. When the price of the alternative k belonging to the group g varies, the cross-price elasticities η_{jk} are not identical whether j belongs to the same group or not. The price elasticities of the demand of the alternative j with respect to the alternative k is:

$$\eta_{jk} = \begin{cases} \frac{-\alpha}{1-\sigma_g} p_j (1 - (1-\sigma_g)s_j - \sigma_g s_{j/g}) \text{ if } j = k \text{ and } j,k \in g \\ \frac{\alpha}{1-\sigma_g} p_j ((1-\sigma_g)s_j + \sigma_g s_{j/g}) \text{ if } j \neq k \text{ and } j,k \in g \\ -\alpha p_k s_k \text{ if } j \neq k \text{ and } j \in g, k \in h \end{cases} . \quad (7)$$

In this discussion, we considered the simple case of the nested logit models with two nests. In some situations, three or more nests may be appropriate, where the probability expression is a relatively straightforward generalization of the two-nest case. Goldberg (1995) considers a five-nest case, while Brenkers and Verboven (2006) use three levels. In general, the parameters of the multinomial and nested logit models can be estimated by maximum likelihood, whereas mixed logit models require the use of the simulated maximum likelihood (Train 2003). Most econometric software (Stata, Nlogit, R) contain algorithms that allow for relatively simple, efficient estimation of any logit variant, whether with random coefficients or not.

3 Models of Discrete-Continuous Choice

For many products – consumer non-durables such as food and beverages or environmental amenities such as parks or fisheries – the choice process is more appropriately described as discrete-continuous than either purely discrete or entirely continuous. There are many classes of goods for which people do not purchase a single-item, but rather a variable amount or variable weight of a specific product. For example, meat, fresh produce, or even bottled water can all be described as discrete-continuous. In this section, we will consider two modeling approaches, and develop one in more detail.

3.1 Discrete-Continuous Choice Models

Why might choices be made in a discrete-continuous way, and what does this imply about the nature of the underlying utility functions? Discrete-continuous choices are typically characteristic of the product-class. Durable goods with either variable amounts of usage or inputs, non-durable goods that are purchased in varying quantities or many services are good examples. The consumer may purchase one alternative out of many in the consideration set, but then purchase an amount that varies continuously, and in a way that differs from other consumers in the data set.

Many choices involve a discrete and then a continuous choice in the same purchase that invalidates the underlying econometric assumptions of our traditional demand model: The choice of a brand or variety and the volume to buy is the most obvious in a food-demand context (Chintagunta 1993). In each case, the relevant data contains a large number of zeros for the alternatives that were not purchased, and continuous purchase amounts for those that were. There are two ways of dealing with this issue econometrically: (1) creating an *ad hoc* econometric model that accounts for the selection bias created by the discrete choice process within a continuous modeling framework, or (2) estimating a model of discrete-continuous choice that is grounded in a single, unifying utility-maximization framework (Wales and Woodland 1983).

The early models of Heckman (1979) and Lee, Maddala and Trost (1980) are of the first form, based indirectly in the theory of utility-maximization theory, but dealing with the econometric issues associated with a censored dependent variable in a statistically-correct way. That is, if the dependent variable is inherently "zero-positive" then there are clear statistical problems with applying standard ordinary least squares estimation methods. The most common method of estimating these models relies on the Heckman two-stage approach in which a probit model is applied to the buy / no-buy problem in the first stage and then the inverse Mill's Ratio is used as a regressor in the second stage OLS regression to correct for the sample selection bias. If there is a way to describe the data generating process directly,

however, it will nearly always be preferred.

Corner solutions to utility-maximization problems can be modeled in an empirically-tractable way. Hanemann (1984) describes an approach based on the Kuhn-Tucker conditions for utility maximization that formally introduces the empirical restrictions on a model of discrete-continuous choice implied by the theory. His approach is similar to that of Wales and Woodland (1983), who describe a method of estimating demand systems in the presence of corner solutions, or zeros in the dependent variable. Hanemann's (1984) model describes a particular setting in which only one choice is made and a continuous amount is purchased. Although this may be a simplification of many choice environments, his approach represented a substantial advance in structural demand modeling.

The intuition behind the approach is as follows: Assume a perfect-substitutes world in which the good with the lowest price-per-unit of quality is purchased (Deaton and Muellbauer 1980). The choice of a particular good is determined in a random utility framework, so is governed by the distribution of the unobserved heterogeneity that drives the specification for perceived quality. Conditional on this choice, therefore, the expected purchase quantity is found by solving for the implied demand from a known indirect utility function, and applying a change of variables from the random unobserved heterogeneity term to the quantity-demand term. The result is an expected expenditure amount that is a parametric function of the arguments of the implicit quality function of the good in question.

Formally, the model consists of two stages, the first describing the discrete choice of goods, and the second the distribution of the continuous volume purchased. The direct utility function for the perfect substitutes model is given by the general class of utility function written as:

$$u(x_1, x_2, \psi_1, \psi_2, z) = u(x_1\psi_1, x_2\psi_2, z), \tag{8}$$

for two goods, where x_j is the quantity of good j, ψ_j is the quality of good j, and z is all other goods such that income, y, is exhausted. This is a perfect substitutes model because maximizing utility subject to an income constraint implies that only one good is purchased,

the one with the lowest ratio of price-to-quality, or p_j/ψ_j. Clearly, the specification for perceived, or expected, quality is key to the model, because it determines which good is purchased. Hanemann (1984) describes both a linear and multiplicative quality function, but we will focus on the multiplicative function with attributes for good j given by b_j so that the quality function is written: $\psi_j(b_j, \varepsilon_j) = \exp(\alpha_j + \sum_k \gamma_k b_{jk} + \varepsilon_j)$,where ε_j is the random component of quality that is assumed to be iid extreme-value distributed in the model development to follow. By parameterizing the quality function this way, the choice probabilities are given by:

$$\pi_j = \Pr[\varepsilon_j + \alpha_j + \sum_k \gamma_k b_{jk} - \ln p_j \geqslant \varepsilon_i + \alpha_i + \sum_k \gamma_k b_{ik} - \ln p_i, \ \forall i, \tag{9}$$

so with the extreme-value assumption the probability of choosing item j becomes:

$$\pi_j = \frac{\exp(\alpha_j + \sum_k \gamma_k b_{jk} - 1/\mu \ln p_j)}{\sum_i \exp(\alpha_i + \sum_k \gamma_k b_{ik} - 1/\mu \ln p_i)}, \tag{10}$$

where μ is the logit scale parameter.

From the logit choice probability, we then find the distribution of demand for the commodity by applying a change of variable technique based on the conditional demand function for x_j. Assuming an indirect utility function from a simple bivariate utility model: $v(p_j, y) = (\theta/(\rho - 1))p_j^{1-\rho} - \exp(-\eta y)/\eta, \ \theta > 0$, then the conditional demand function is found by applying Roy's theorem to find: $x_j(p_j, \psi_j, y) = \theta p_j^{-\rho} \psi_j^{\rho-1} \exp(\eta y)$, and, after substituting the expression for quality:

$$x_j(p_j, \psi_j, y) = \theta p_j^{-\rho} \exp(\eta y) \exp((\rho - 1)\lambda_j) \exp((\rho - 1)\varepsilon_j), \tag{11}$$

where $\lambda_j = \alpha_j + \sum_k \gamma_k b_{jk} - \ln p_j$, or the mean quality function less prices. We then apply a change of variables from ε_j to x_j and take the mean of the resulting conditional distribution to find:

$$E[\ln p_j x_j | \varepsilon_j + \lambda_j > \varepsilon_i + \lambda_i] = \ln \theta + \eta y + (\rho - 1)[\mu \ln(\sum \exp(\lambda_i/\mu)) + 0.5772\mu], \tag{12}$$

where 0.5572 comes from the expectation of an EV random variable.

The expected demand function can then be estimated using maximum likelihood, or the two-stage estimator described in Hanemann (1984). The two-stage estimator involves estimating the values of λ_j/μ from a logit maximum likelihood routine, and then using the estimated values in the demand equation to estimate the remaining parameters with OLS. Although this was the recommended approach in 1984, it is more efficient to estimate everything together with MLE. Because this discrete-continuous specification is derived from a single utility maximization problem, the choice to purchase and how much to purchase are internally consistent, but the primary drawback is that price elasticities are restricted to -1. Despite this fact, the model is relatively flexible as Hanemann (1984) describes several other utility specifications that will work in this framework. Applications of the multiple-discrete model to estimating food demand include Chintagunta (1993) and Richards (2000).

4 Models of Multiple-Discrete and Multiple-Discrete Continuous Choice

In the last two decades, researchers in transportation (Bhat 2005, 2008), marketing (Hendel 1999; Dube 2004) and environmental economics (Phaneuf et al. 2000) recognized that individuals in many settings not only make discrete-continuous choices, but often make multiple discrete choices, such as choosing several brands of soda on each trip to the store, or more than one variety of apple. In this section, we describe three models that are able to address the: (a) multiple-discrete, (b) multiple-discrete continuous, and (c) multiple-discrete continuous with complementarity issues in flexible, tractable ways. We introduce the intuition underlying the first two specifications and develop the latter, most general model, more formally.

4.1 Models of Multiple-Discrete Choice

Like the models of discrete-continuous choice developed above, models of multiple-discrete-continuous (MDC) choice are also grounded in the theory of utility-maximization. However, they tend to be comprised of sub-problems, each describing a different part of the decision process, that are solved together in one utility-maximization framework. MDC models written this way explain an important observation in quantitative marketing, namely if consumers make multiple, discrete purchases on each trip to the store, then there is a revealed demand for variety. For example, if a consumer buys both Diet Coke and Coke on a trip to the store, they are clearly either anticipating a change in tastes from consumption occasion to consumption occasion (between purchase occasions), or are buying for others in the family. The structure of the model that accounts for this demand for variety is based on the general theme of identifying corner solutions from a single utility-maximization problem (Dube 2004).

The utility maximization process assumes consumers have a number of consumption occasions between purchases. The total utility from a purchase occasion, therefore, sums over the sub-utility functions that describe the utility from each consumption occasion. Consumers maximize the utility from a purchase occasion and not a consumption occasion. Therefore, the expected quantity purchased at each visit to the store is composed of the distribution of demands for each consumption occasion and the distribution that governs the number of consumption occasions (a count-distribution). The three components to the demand model are, therefore, (1) the count-data model that governs the number of consumption occasions, (2) the sub-utility function that determines what is consumed on each consumption occasion, and (3) the total utility maximization process at each purchase occasion. Consumers are assumed to maximize utility subject to a budget constraint, and the Kuhn-Tucker conditions are used to derive estimable demand models for each purchase occasion. MDC models are able to produce elasticities that appear reasonable, and have proven useful in applied industrial organization models, where accurately conditioning for consumer demand is critical.

4.2 Models of Multiple-Discrete Continuous Choice

The MDC model described above, however, assumes that each of the purchases is still only discrete, and that consumers either purchase a constant amount, or the discrete purchases themselves are for different quantities. In this section, we describe a model that synthesizes the corner-solution approach developed in Section 2, with the multiple discrete logic outlined above. Originally applied to problems in transportation (Bhat 2005, 2008), where individuals often choose multiple modes of transportation, and use each for varying distances or amounts of time, the application to food demand is fairly obvious. Namely, for many categories of products, consumers purchase many different brands, or varieties, in the same category, and purchase a continuous amount of each. For example, Richards, Gomez, and Pofahl (2011) describe a problem in the demand for fresh produce. Items within each sub-category, apples for example, are purchased by the variety, but the amounts are typically measured in pounds. A substantial proportion of the consumers in that data reported purchasing multiple varieties on each purchase occasion, whether due to varying tastes within the household, or a desire to not have to eat the same kind of apple time after time.

As in the MDC case, the underlying model is consistent with utility maximization, and the Kuhn-Tucker conditions for constrained utility maximization are used to derive the demand model. Unlike the MDC model, however, the multiple-discrete continuous model of Kim, Allenby, and Rossi (2002) and Bhat (2005, 2008) generates demand equations that describe the joint probability distribution for continuous quantities of a discrete set of items chosen from a larger consideration set. By including the utility from a numeraire good that is always consumed, demands for each of the other "inside goods" are derived using an equilibrium argument: The utility from a good that is purchased must be at least as great as the utility from the always-consumed numeraire good. Assuming consumers make random errors in utility maximization, and that these errors are Type I Extreme Value distributed, the resulting system of purchase probabilities is derived. Remarkably, the demand equations nest a simple logit model when only one item is purchased. Bhat (2005, 2008) shows how

unobserved heterogeneity can be accommodated by allowing for random parameters in the usual way. Typically, the resulting multiple-discrete continuous extreme value (MDCEV) model is estimated using simulated maximum likelihood.

4.3 Generalized Model of Multiple-Discrete Continuous Choice

The MDCEV model described above has become a common method of estimating multiple-discrete continuous demand models. However, this class of model still retains a critical weakness that all products are restricted to be substitutes. When the problem involves items in multiple categories – milk, bread, and cereal, for example – then any reasonable model would need to accommodate the possibility that some items may be complements. Pinjari, Castro and Bhat (2012) and Vasquez-Lavin and Hanemann (2008) derive generalized versions of the multiple-discrete continuous model described above that does just that. Pairs of items can be complements, depending on the sign of an interaction parameter. Formally, the utility function for this Generalized Multiple Discrete Continuous Extreme Value (GMDCEV) model is written as:

$$u_j^h(q_{ij}^h,\Omega) = \frac{1}{\alpha_1}(q_{1j}^h)^{\alpha_1}\phi_{1j}^h + \sum_{i=2}^{I}\left[\frac{\gamma_i}{\alpha_i}\left(\left(\frac{q_{ij}^h}{\gamma_i}+1\right)^{\alpha_i}-1\right)\left(\phi_{ij}^h + 1/2\sum_{i\neq k, i\neq 1}^{K}\theta_{ik}\frac{\gamma_k}{\alpha_k}\left(\left(\frac{q_{kj}^h}{\gamma_k}+1\right)^{\alpha_k}-1\right)\right)\right], \tag{13}$$

$j = 1, 2...J, \quad h = 1, 2, ...H$,where q_{ij}^h is the amount of good i purchased by household h on occasion j, Ω is a vector of parameters to be estimated,

$$\pi_{ij}^h = \phi_{ij}^h + 1/2\sum_{i\neq k, i\neq 1}^{K}\theta_{ik}\frac{\gamma_k}{\alpha_k}\left(\left(\frac{q_{kj}^h}{\gamma_k}+1\right)^{\alpha_k}-1\right) \tag{14}$$

is the baseline marginal utility, for good i on occasion j by household h ($\pi_{ij}^h > 0$), α_i are parameters that reflect the curvature of the utility function ($-\infty < \alpha_i \leq 1$) and γ_i is the product-specific utility translation parameter ($\gamma_i > 0$). Note that, because $\gamma_1 = 0$ by assumption, the numeraire good is not subject to satiation effects.

The parameters α_i and γ_i are largely what separate the MDCEV (GMDCEV in our case) model from others in the class of discrete, multiple-discrete, or discrete-continuous models

(Richards, Gomez, and Pofahl, 2012). In mathematical terms, γ_i is a translation parameter that determines where the indifference curve between q_{1j} and q_{2j} becomes asymptotic to the q_{1j} or q_{2j} axis, and thereby where the indifference curve intersects the axes. The parameter α_i, on the other hand, determines how the marginal utility of good i changes as q_{ij} rises. If $\alpha_i = 1$, then the marginal utility of i is constant, indifference curves are linear, and the consumer allocates all income to the good with the lowest quality-adjusted price (Deaton and Muellbauer, 1980). As the value of α_i falls, satiation rises, the utility function in good i becomes more concave, and satiation occurs at a lower value of q_{ij}. Importantly, if the values of π_{ij}^h are approximately equal across all varieties, and if the individual has relatively low values of α_i, then he or she can be described as "variety seeking" and purchase some of all choices, while the opposite will be the case if α_i are relatively high (close to 1.0) and the perceived qualities differ (Bhat 2005).

The GMDCEV incorporates additive separability in a form suggested by Vasquez-Lavin and Hanemann (2008) in that utility is quadratic in quantities. Both complementary ($\theta_{ik} > 0$) and substitute ($\theta_{ik} < 0$) relationships are permitted between pairs of products, so the GMDCEV represents a very general corner-solution model. Bhat, Castro and Pinjari (2012) show that allowing unrestricted own-quadratic effects can lead to negative values for baseline utility, so restrict $\theta_{ii} = 0$. This restriction makes sense as the data are not likely to identify additional non-linearities with respect to own-price effects, but should reflect interactions between products that are not part of any additively-separable demand system. Each of the constant terms, or ϕ_{ij}^h parameters can also be written as functions of demographic or marketing mix variables to address concerns regarding the importance of observed heterogeneity.

As with the MDC and MDCEV models, the Kuhn-Tucker approach is used to solve for discrete / continuous demand system implied by the utility function described above. By solving the Kuhn-Tucker conditions for the constrained utility maximization problem, the GMDCEV demand functions consist of a mixture of corner and interior solutions that are a product of the underlying utility structure, and are not simply imposed during econometric

estimation. The details of the solution procedure are in Pinjari, Castro, and Bhat (2012), so we will only summarize the resulting demand system here.

Assuming the optimization procedure is solved only up to a random error ε_{ij}^h, and assuming the errors are distributed Type I Extreme Value, the econometric model assumes a particularly straightforward form as the utility-maximizing solution collapses to:

$$P(q_{1j}^h, q_{2j}^h, ..., q_{Mj}^h, 0, 0...0) = \frac{1}{\mu^{M-1}}|J|(\prod_{k=1}^{M} e^{V_{kj}^h/\mu})(\sum_{i=1}^{I} e^{V_{ij}^h/\mu})^{-M}(M-1)!, \tag{15}$$

where $|J|$ is the Jacobian of the transformation from the errors to the demand quantities, $V_{kj}^h(p_1, p_2, ..., p_I, y^h)$ is the indirect utility function implied by the choice model above, and M varieties are chosen out of I available choices. In this estimating equation, μ is the logit scale parameter. In fact, when $M = 1$, or only one alternative is purchased, and there is no cross-category effects, the GMDCEV model becomes a simple logit. Although appearing complicated, the GMDCEV model is estimated in a straightforward way using MLE, or in a random-coefficient variant using the SML approach described in the final section below.

5 Shopping-Basket Models

Consumers typically purchase many items together, from dozens of categories, and many brands. Typically, empirical models of consumer demand focus only on one category at a time, ignoring potentially-important interactions with items in other categories. If consumers purchase groceries by the shopping basket, and not just one item at a time, then it is reasonable to estimate models that take into account the demand for many categories, and potential for complementarity, on each shopping occasion (Ainslie and Rossi 1998; Manchanda, Ansari, and Gupta 1999; Russell and Petersen 2000; Chib, Seetharaman, and Strijnev 2002; Kwak, Duvvuri, and Russell 2015). Complementarity matters, because retailers set prices as if consumers purchase items together, in the same shopping basket (Smith 2004). In this section we present an alternative way of modeling consumers' shopping-basket choice process: The multivariate logit (MVL) model.

5.1 Model of Retail Demand

Like the GMDCEV model, the MVL model is derived from a single utility-maximization process. Unlike the GMDCEV, however, the choice process is based on the random-utility assumption. We begin by describing the nature of the MVL utility function, and how it is used to describe consumers' choices among several items that may appear together in their shopping basket. Consumers $h = 1, 2, 3, ..., H$ in the MVL model select items from among $i = 1, 2, 3, ..., N$ categories, c_{iht}, in assembling a shopping basket, $\mathbf{b}_{ht} = (c_{1ht}, c_{2ht}, c_{3ht}, ..., c_{Nht})$ on each trip, t, conditional on their choice of store, r. Define the set of all possible baskets in r as $\mathbf{b}^r_{ht} \in \mathbf{B}^r$. Our focus is on purchase incidence, which is the probability of choosing an item from a particular category on a given shopping occasion, and we model demand at the category level by assuming consumers purchase one item per category across multiple categories.

Consumers choose among categories to maximize utility, U^r_{ht}, and we follow Song and Chintagunta (2006) in writing utility in terms of a discrete, second-order Taylor series approximation:

$$\begin{aligned} U^r_{ht}(\mathbf{b}^r_{ht}|r) &= V^r_{ht}(\mathbf{b}^r_{ht}|r) + \varepsilon^r_{ht} \\ &= \sum_{i=1}^{N} \pi^r_{iht} c^r_{iht} + \sum_{i=1}^{N} \sum_{j \neq i}^{N} \theta^r_{ijh} c^r_{iht} c^r_{jht} + \varepsilon^r_{ht}, \end{aligned} \tag{16}$$

where π^r_{iht} is the baseline utility for category i earned by household h on shopping trip t in store r, c^r_{iht} is a discrete indicator that equals 1 when category i is purchased in store r, and 0 otherwise, ε^r_{ht} is a Gumbel-distributed error term that is iid across households and shopping trips, and θ^r_{ijh} is a household-specific parameter that captures the degree of interdependence in demand between categories i and j in store r. Specifically, $\theta^r_{ijh} < 0$ if the categories are substitutes, $\theta^r_{ijh} > 0$ if the categories are complements, and $\theta^r_{ijh} = 0$ if the categories are independent in demand. To ensure identification, we restrict all $\theta^r_{ii} = 0$ and impose symmetry on the matrix of cross-purchase effects, $\theta^r_{ijh} = \theta^r_{jih}, \forall i, j \in r$ (Besag 1974, Cressie 1993, Russell and Petersen 2000).

The probability that a household purchases a product from a given category on a given shopping occasion depends on both perceived need, and marketing activities from the brands in the category (Bucklin and Lattin 1992, Russell and Petersen 2000). Therefore, the baseline utility for each category depends on a set of category ($\mathbf{X}_i$) and household ($\mathbf{Z}_h$) specific factors such that: $\pi^r_{iht} = \alpha^r_{ih} + \beta^r_{ih}\mathbf{X}^r_i + \gamma^r_{ih}\mathbf{Z}_h$,where perceived need, in turn, is affected by the rate at which a household consumes products in the category, the frequency that they tend to purchase in the category, and any other household demographic measures. Category factors include marketing mix elements, such as prices, promotion, or featuring-activities. As with any other demand model, unobserved heterogeneity can be included by allowing any of these parameters to be randomly distributed over households.

With the error assumption in equation (16), the conditional probability of purchasing in each category assumes a relatively simple logit form. Following Kwak, Duvvuri, and Russell (2015), we simplify the expression for the conditional incidence probability by writing the cross-category purchase effect in matrix form, suppressing the store index on the individual elements, where: $\Theta^r_h = [\Theta_{1h}, \Theta_{2h}, ..., \Theta_{Nh}]$ and each Θ_{ih} represents a column vector of a $N \times N$ cross-effect Θ^r_h matrix with elements $\theta^r_{ijh.}$. With this matrix, the conditional utility of purchasing in category i is written as:

$$U^r_{ht}(c^r_{iht}|c^r_{jht}) = \pi^{r\prime}_{ht}\mathbf{b}^r_{ht} + \Theta^{r\prime}_{ih}\mathbf{b}^r_{ht} + \varepsilon_{ht}, \tag{17}$$

for the items in the basket vector $\mathbf{b}^r_{ht}$. Conditional utility functions of this type potentially convey important information, and are more empirically tractable that the full probability distribution of all potential assortments (Moon and Russell 2008), but are limited in that they cannot describe the entire matrix of substitute relationships in a consistent way, and are not econometrically efficient in that they fail to exploit the cross-equation relationships implied by the utility maximization problem. Estimating all N of these equations together in a system is one option, but Besag (1974) describes how the full distribution of $\mathbf{b}^r_{ht}$ choices are estimated together.

Assuming the Θ^r_h matrix is fully symmetric, and the main diagonal consists entirely of

zeros, then Besag (1974) shows that the probability of choosing the entire vector $\mathbf{b}_{ht}^r$ is written as:

$$\Pr(\mathbf{b}_{ht}^r|r) = \frac{\exp(\boldsymbol{\pi}_{ht}^{r\prime}\mathbf{b}_{ht}^r + \frac{1}{2}\mathbf{b}_{nt}^{r\prime}\Theta_h^r\mathbf{b}_{ht}^r)}{\sum_{\mathbf{b}_{ht}^r \in \mathbf{B}^r}[\exp(\boldsymbol{\pi}_{ht}^{r\prime}\mathbf{b}_{ht}^r + \frac{1}{2}\mathbf{b}_{ht}^{r\prime}\Theta_h^r\mathbf{b}_{ht}^r)]}, \tag{18}$$

where $\Pr(\mathbf{b}_{ht}^r)$ is interpreted as the joint probability of choosing the observed combination of categories from among the 2^N potentially available from N categories, still conditional on the choice of store r. Assuming the elements of the main diagonal of Θ^r is necessary for identification, while the symmetry assumption is required to ensure that (18) truly represents a joint distribution, a multi-variate logistic (MVL, Cox 1972) distribution, of the category-purchase events. Essentially, the model in (18) represents the probability of observing the simultaneous occurrence of N discrete events – a shopping basket – at one point in time.

Given the similarity of the choice probabilities to logit-choice probabilities, the elasticities are similar to the those shown above for the logit model, but recognizing the fact that cross-price elasticities for items within the same basket will differ from those in different baskets (Kwak, Duvvuri, and Russell 2015). In the absence of unobserved heterogeneity, the MVL model is estimated using maximum likelihood in a relatively standard way, but when random parameters are used, the model is estimated using the SML method described below.

The MVL is powerful in its ability to estimate both substitute and complementary relationships in a relatively parsimonious way, but suffers from the curse of dimensionality. That is, with N products, the number of baskets is $N^2 - 1$, so the problem quickly becomes intractable for anything more then a highly stylized description of the typical shopping basket.

6 Spatial Econometrics and the Distance Metric Model

There is a rich history of modeling the demand for differentiated products solely in terms of their attributes (Lancaster 1966). In fact, the mixed logit model relies on attribute variation among items in a category of products to identify differences in price elasticities,

and to project the demand for differentiated products from a high-dimensional product space, to a lower-dimensional attribute space. It is both convenient and intuitive to think of products not necessarily in terms of their brand or variety, but in terms of the attributes that comprise them. Slade (2004) exploits attribute variation among a large number of beer brands in developing the "distance metric" (DM) demand model as an alternative means of overcoming the curse of dimensionality in differentiated-products demand analysis, and avoiding the IIA problem associated with logit models. In this section, we briefly review the power of spatial econometrics more generally, and show how the DM model represents a fundamentally-different way of estimating demand.

6.1 Spatial Econometrics and Demand Estimation

In this model, attribute-variation is another way of circumventing the IIA characteristic of logit-based demand systems. Because the distance between products in attribute space as a primitive of the consumer choice process, the matrix of substitution elasticities is completely flexible, unlike a simple logit. Slade (2004) applies a similar notion of product differentiation to the discrete choice model by assuming the price-coefficient to be a function of attributes; however, a disadvantage of this approach is that a consumer's price-response in a discrete-choice model of demand is determined by the marginal utility of income, which is a characteristic of the individual that cannot logically vary over choices. Rather, the DM model described here includes attribute-distance as a direct argument of the utility function.

The DM approach to demand estimation is similar to the address model of Anderson, de Palma, and Thisse (1992) and Feenstra and Levinsohn (1995) in that the utility from each choice depends upon the distance between the attributes contained in that choice and the consumer's "ideal" set of product attributes, where the ideal product reduces to the product chosen by a representative consumer. The DM models accounts for the utility-loss associated with distance in by introducing a spatial autoregression parameter to measure the extent to which differentiation from other products raises (or lowers) the utility from choosing product

j according to the relative distances between products and the ideal attribute mix of a given consumer.

The distance metric - multinomial logit (DM-MNL) uses a non-linear utility-loss function, where mean utility from product j falls (or rises) in the distance from all other products, measured by the distance matrix $\mathbf{W}$. Each element of $\mathbf{W}$ measures the Euclidean distance between each pair of product, so the element w_{jl} measures the distance between product j and product l in a multi-attribute space. The importance of differentiation is estimated through a spatial-autoregressive parameter. Formally, mean utility for product $j = 1, 2, ..., J$ in week $t = 1, 2, ..., T$ is written in vector notation (with bold notation indicating a vector) as:

$$\boldsymbol{\delta} = \boldsymbol{\beta}'\mathbf{x} + \lambda\mathbf{W}\boldsymbol{\delta} - \alpha\mathbf{p} + \boldsymbol{\xi}, \tag{19}$$

where $\boldsymbol{\delta}$ is a $JT \times 1$ vector of mean utility, $\mathbf{x}$ is a $JT \times K$ matrix of demand shifters, $\mathbf{p}$ is a $JT \times 1$ vector of prices, and $\boldsymbol{\xi}$ is a random error unobserved by the econometrician. The vector $\boldsymbol{\beta}$ and scalar parameters λ and α are all estimated from the data. The matrix $\mathbf{W}\boldsymbol{\delta}$ measures the effect of product differentiation on utility according to attribute distance, which defines λ as a spatial autoregression parameter (Anselin, 2002).

As a spatial autoregression parameter, λ is interpreted as the extent to which utility is affected, positively or negatively, by the distance between the chosen product, and all other products in the choice set. Autoregression reflects the notion that consumers evaluate the utility attainable from each product relative to the utility that can be attained from consuming other available products in the choice set. By convention, $\mathbf{W}$ is defined as a measure of inverse-distance, or proximity, so that greater product differentiation in the product category reduces utility when $\lambda > 0$ (i.e., utility rises with attribute proximity) and increases utility when $\lambda < 0$.

Solving equation (19) for mean utility gives: $\boldsymbol{\delta} = (\mathbf{I} - \lambda\mathbf{W})^{-1}(\boldsymbol{\beta}'\mathbf{x} - \alpha\mathbf{p} + \boldsymbol{\xi})$,where $(\mathbf{I} - \lambda\mathbf{W})^{-1}$ is the Leontief inverse, or spatial multiplier matrix (Anselin, 2002). In spatial models, the concept of the multiplier is critical, and powerful, because it measures how changes to

one observation ripple throughout the entire system. For example, if one price changes exogenously, the demand for all other products changes according to the spatial multiplier matrix. In a context where $\mathbf{W}$ measures the distance between individuals consuming the product in a social network, the multiplier measures how strong the peer- or bandwagon-effects for the product are.

Assuming utility varies among consumers in a random way, utility is written as: $\mathbf{u}_i = \boldsymbol{\delta} + \boldsymbol{\varepsilon}_i$,where $\boldsymbol{\varepsilon}_i$ is an iid random error that accounts for unobserved consumer heterogeneity. Further assuming ε_{ij} is Type I Extreme Value distributed, and aggregating over consumers, the DM-MNL model yields a market share expression for item j given by: $S_j = \exp(\delta_j)/(1 + \sum_{l=1}^{J} \exp(\delta_l))$,where S_j is the volume-share of product j, which can be linearized using the approach in Berry (1994) and Cardell (1997) and estimated using MLE. However, because the $\mathbf{W}$ matrix must be inverted during estimation, a MLE routine may encounter computational issues. Kelejian and Prucha (1999) describe a generalized method of moments (GMM) routine that avoids these issues, and accounts for the likely endogeneity of prices, or any other marketing mix elements for that matter.

There are many other ways of applying the DM concept to demand modeling. The MNL model above is similar to Slade (2004) and Pinkse and Slade (2004) in that we explicitly incorporate a distance-metric component in the demand model; however, attribute distance enters in a structural way in equation (19) through the utility function. Rojas and Peterson (2008) and Pofahl and Richards (2009) describe two other approaches using more traditional demand systems. The point is that including attribute space through the DM logic is very general – projecting demand into attribute space, or even social space (Richards and Hamilton 2014) not only reduces the dimensionality problem associated with differentiated-products analysis, but adds flexibility and the ability to study a wider range of applied problems.

7 Machine Learning

Advances in computing power, and in the creation of huge data sets generated by virtually any web-based activity, have renewed interest in "big data" methods for analyzing consumer-demand problems (Varian 2014; Bajari et al. 2014). While the definition of what exactly constitutes big data remains elusive, it has come to be associated instead with a set of analytical methods rather than attributes of the data itself. When presented with virtually unlimited numbers of observations, and possibly thousands of explanatory variables, researchers have turned to machine learning (ML) methods rather than traditional econometric techniques. Using ML methods to analyze demand data, however, is fundamentally different from any of the frameworks discussed above in that the outputs are different, and the objectives of the analysis differ accordingly.

7.1 Studying Demand Data with ML Methods

ML, or statistical learning more generally, is typically used as a prediction tool. In fact, models are evaluated on the basis of their ability to fit out-of-sample, instead of on some sort of in-sample metric as is usually the case in econometrics. The model that is able to produce the lowest root mean squared error (RMSE) on a cross-validation sample of the data is the winner. That said, recent advances in the literature on machine learning investigate how big data models can be used to study causal inference (Athey and Imbens 2015) or to generate marginal effects similar to econometric models of demand (Bajari et al. 2014; Varian 2014). In this section, we will review 6 machine learning techniques, and how they can be applied to demand data. Our discussion draws heavily on James et al. (2014), which is a valuable and standard reference in this area.

Many of the methods are actually variants on standard econometric approaches, using the concept of least squares in different ways to estimate large models. At the risk of over-simplification, these methods can be classified into either *regularization* approaches, or *tree-based* methods. Regularization involves reducing a regression problem to a smaller one

by restricting some coefficients that are close to zero, exactly to zero, focusing on the non-zero estimates. Tree-based methods, on the other hand, seek to order predictor variables according to their importance, and determine critical breaks in regions of statistical support. In this section, we consider three of the former (forward stepwise regression, lasso, and support vector machines) and three of the latter (bagging, random forests, and boosting). We also provide a brief discussion of cross-validation as a method for model selection.

7.2 Regularization and Penalized Regression

Analysts in the ML literature generally have no qualms with using *forward stepwise regression* as a method for selecting the best linear model. Forward stepwise regression begins by estimating a null model, and then adding variables in succession and choosing the predictor at each step that produces the lowest cross-validated prediction error, AIC, BIC, or adjusted R^2. While econometricians may have conceptual issues with the data mining aspect of forward stepwise regression, that is the point of machine learning. With large data sets, of very high dimension, forward stepwise regression is often a very pragmatic, and effective, tool for model selection, particularly given the power of cross-validation when the size of the data set permits holding out a large number of observations for training purposes.

A second class of models is known as shrinkage, penalized regression, or regularization methods. Regularization means that the coefficients on some predictors are reduced to zero in estimation if their statistical effect is, for all practical purposes, zero. They are referred to as shrinkage methods because they effectively shrink the size of the predictor set according to the number of zero coefficients that are assigned. Principal among these methods is the lasso, which minimizes an objective function that includes a penalty for many, large regression coefficients:

$$LASSO = \min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{20}$$

where λ is referred to as a "tuning parameter" that controls the extent to which the choice of parameters is constrained by the penalty. When $\lambda = 0$, lasso estimates are clearly equal to

the least squares estimates, and when λ is sufficiently large, all parameter estimates will be reduced to zero. Lasso estimates are particularly valuable in settings where p is large relative to n – that is, in high-dimensional data sets with relatively few observations. In this case, the approach has the effect of shrinking parameter estimates for non-important variables to zero, effectively becoming a means of selecting variables based on their values as predictors. With sufficient data, cross-validation methods over a grid that includes a wide range of possible parameter values is used to determine the value of λ that minimizes out-of-sample forecast error. As a shrinkage model, lasso is similar to ridge regression, but the latter, which uses a quadratic rather than absolute-value penalty, never reduces any coefficient estimates exactly to zero, but only shrinks them toward zero. If the problem is dimensionality, ruling some variables out is important.

Support vector machines (SVM) are designed for classification, that is, assigning observations in the data set to binary classes. They are unique in that they rely on the notion of a *maximal margin classifier* (MMC) which is an algorithm that chooses the parameters of a separating hyperplane – familiar to economists as the core construct in duality theory – in order to maximize the minimum distance between the hyperplane and data observations. However, the base MMC method suffers from the fact that the data are often not sufficiently well behaved to identify a unique hyperplane that cleanly separates all the observations into one class or another. That is, the MMC solution does not exist.

Consequently, the SVM approach is based on a *support vector classifier* (SVC) method that allows for some observations to lie on the wrong side of the margin, or even on the wrong side of the hyperplane. In the SVC optimization routine, however, only observations that either lie on or on the wrong side of the margin enter into the calculation, as the objective function values for the others are very small. Therefore, these vectors are known as *support vectors* as they determine the location of the margin alone. Despite the fact that the SVC method is more flexible than the MMC in the sense that it admits violations of the strict MMC principle, it still constrains the margin to be linear. In many, if not most, data sets, the

classification margin is not linear. SVM were developed specifically to allow for non-linear classification margins.

Support vector machines (SVM) are a special class of SVC that introduce a larger feature space created from polynomials of the original features. Although a margin defined with a SVM is still linear in the expanded set of features, it can be highly non-linear in the original, un-transformed features. The SVM algorithm is the same as that developed for the SVC, but relies on the recognition that only the support vectors matter. That is, the others that do not enter the solution are formally excluded. And, the calculation used to find the location of the margin depends only on the inner product of all the vectors that matter, or the *kernel* of the data. When the kernel is linear, the inner product is simply the correlation between each pair of vectors. But, different kernels can be used to allow for support vectors that describe highly non-linear class boundaries. For example, a polynomial kernel of degree d can produce non-linear boundaries, and a radial kernel even describes a circular region of support, separating observations into highly flexible patterns of association within the data. In essence, a SVM is a SVC with a non-linear kernel.

7.3 Tree-Based Methods

Regression trees, on the other hand, are a means of determining the relative importance of a predictor variable in influencing an output variable. If the data are continuous, a regression tree algorithm searches for a split value of the most important predictor, and then calculates predicted values for the output variable for values above and below the split value. Once all observations are assigned in one branch, the algorithm then seeks the predictor variable that best explains the next split for each of the new branches, and so on. Because this *recursive binary splitting* algorithm begins at the top and makes the error-minimizing decision for that split only, it is referred to as a greedy algorithm (James et al. 2014).

Predictive accuracy is evaluated out-of-sample through a k-fold cross validation method: Divide the training data into $k = 1, 2, ..., K$ subsets, or folds (of equal size), train the model

on the data in $k-1$ folds, and calculate the mean-square-error (MSE) on the kth fold. Repeat for each of the other k-folds, estimating on each of the other k and finding MSE on the $k-1$ fold so that there are k estimates of the MSE, and average the MSE that results. The result is a measure of the k-fold cross-validated MSE. A simpler alternative is leave-one-out cross-validation (LOOCV), which excludes one observation from training, and then fits the model on the left-out data. However, the LOOCV measure has high variance as the fitted value is averaged over only one observation per run.

Formally, the objective function for a standard, regression tree approach minimizes the residual sum of squares (RSS) given by: $RT = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$,where y_i is the observed value of the variable of interest, and $\hat{y}_{R_j}$is the mean value of the variable in the region R_j. In words, the tree structure divides the data into regions based on values of the predictor space $X_1, X_2, ..., X_p$ and then calculates mean values of the response variable for each realized value of the predictor variables, and chooses the regions in order to minimize the residual sum of squares. James, et al. (2014), however, argue that the base regression tree approach may not produce the best result. Other approaches that average predictions over many trees – a forest of them, in fact – can typically outperform classical methods of classification or prediction.

The three most common methods are bagging, random forests, and boosting. Bagging, or bootstrap aggregation (Breiman 1996), draws a large number of random samples from the data (bootstrap samples), and fits regression trees using cross-validation to determine the optimal structure of each tree. By averaging the predictions from all the bagged predictions, a sum-of-squares minimizing prediction set is derived. Bagging typically represents a substantial improvement in predictive ability relative to a basic regression tree because averaging over a large number of samples provides much more information than a simple, single sample. Intuitively, when the metric for refining the fit of the tree is cross-validation, averaging across different slices of the same data set is far more likely to produce results that are representative of the data generating process as a whole. Bagging suffers, however, when

one or two predictors dominate so that each random sample produces a tree that looks like all the others.

Random Forests represents a variant on the bagging approach in which a random sample of m predictors out of the total set of p predictors is considered at each split in the tree. Only one of the m predictors can be used at each split, and a new sample is drawn each time a split decision is to be made. In this way, some models will contain entirely different predictors than others as not every model can simply draw on the most important predictor every time. When bagged regression-tree models are not constrained in this way, their predictions will be highly correlated, so averaging the predictions does not produce much benefit as each separate run does not add much new information. In fact, bagging is a special case of the random forests method as bagging and random forests are exactly equivalent when the number of predictors in the random forest algorithm (m) is set equal to the total number of predictors (p). By "de-correlating" the predictions from the models, and then averaging, the result typically produces more accurate predictions because each new sub-sample brings independent information to finding which variable is most important in predicting values of the variable of interest (James et al. 2014). In general, the number of predictors in each sample is set at fraction (1/3) of the total number of predictors. Comparing a number of alternative regression tree methods, Bajari, et al. (2014) and Varian (2014) find that the random forests approach is the most effective in minimizing MSE in out-of-bag (OOB) samples.

In a regression tree context, boosting uses the notion of fitting several trees to the same data in a fundamentally different way. Boosting uses a process of "slow learning" in which the tree is not built on many independent bootstrapped samples as in bagging, but in sequence, building on the tree fit before it. Each tree is relatively small, with potentially only a few terminal nodes. Once the initial tree is fit to the training data set, the residuals are saved and a new tree is fit to the residuals. In this way, the boosting algorithm proceeds in a manner that is similar to stepwise regression, considering new predictors in sequence

until the remaining residuals are minimized. At each iteration, or new tree, the updated predictions are only allowed to be influenced by the new predictions up to a "shrinkage parameter," λ, that causes the evolution of the tree to move more slowly. Typically, the λ parameter is set at 0.01 or 0.001. Boosted regression trees that evolve slowly are typically the best performing.

8 Practical Considerations

8.1 Data Sources

Historically, econometricians began studying markets for differentiated products using aggregate datasets. The data consisted of markets shares or volume sold, average prices, and primary product attributes for each product over several time periods and/or geographical areas (Berry, Levinsohn and Pakes 1995). Econometricians interested in food demand are relatively lucky as firms such as Nielsen and IRI Marketing Research began collecting "syndicated" scanner data on a highly disaggregated basis in the late 1990s.[2] Scanner data provides price and movement data on individual items, called Stock Keeping Units (SKU)s, or Universal Product Codes (UPCs). IRI, Nielsen, and Kantar in Europe also maintain consumer panel data sets. Consumer panel data are collected by individual households with hand-held scanning devices. They also contain detailed information on the product, the place of the purchase, and, importantly, attributes of the household. However, household panel data sets do not provide any information about the alternatives that the consumer faces on each shopping occasion.

8.2 Choice Sets

When the set of the alternatives that the consumer faces is not known by the researcher, additional assumptions are needed. Most traditional demand models are estimated under the

[2]Syndication means that cooperating stores send their data to IRI or Nielsen, who then combine the chain-specific data to produce standardized data sets of the entire market, and then share the data with retailers and manufacturers.

assumption that consumers are aware of all available alternatives or use information at the aggregate-level data to infer the set of available alternatives for consumers as Berry, Levinshon and Pakes (1995) or Nevo (2001). However, in markets with rapidly changing product lines or stock-outs, it seems unlikely that consumers have full information on all alternatives. Researchers in marketing and economics highlight how the limited cognitive abilities of consumers restrict their attention to some alternatives (Mehta, Rajiv, and Srinivasan 2003). Hence, the choice set is reasonably assumed to be heterogeneous across consumers, limited in size, and endogenously determined. For example, Bruno and Vilcassim (2008) extend traditional discrete choice models using a random distribution of choice sets and find that not accounting for varying product availability on the UK chocolate confectionery market leads to biased demand estimates. Further, Goeree (2008) estimates a discrete choice model with limited consumer information using advertising data and consumer characteristics and finds that full information models predict upward biased price elasticities that imply greater competition among firms than is realistically the case.

8.3 Outside Good

In order to predict changes in total demand in response to a price change, researchers need to include a measure of how much demand can change, regardless of the set of goods in the choice set. This is accomplished through the outside option. The outside option represents either an aggregate of other alternatives that are considered as further substitutes, or non-purchasing behavior. If the outside option is not included, then the model can be used to predict changes in market shares among consumers who already chose the alternatives, or conditional demand, but not in total demand because the model essentially does not contain any room to expand. In general, for discrete, discrete-continuous or multiple discrete-continuous models, the mean baseline utility for one option is typically set to zero. This definition of the outside option, which amounts to delimiting the relevant market when competitive analysis is the goal, is a key issue as it could affect the level of utility, and subsequent price-elasticity

estimates (Foncel and Ivaldi 2005). In the literature, different approaches have been taken depending on the dataset used. For example, Besanko et al. (1998) use the number of all household shopping trips to compute the share of non-purchase behavior. Berto Villas Boas (2007) restricts her analysis to primary brands and retailers, and then defines the other small brands and retailers as the outside option. Bonnet and Réquillart (2013) use observed purchases of other product categories that are more or less substitutes for their focal soft drink categories to define the outside option. More formally, the relevant market and the outside option could be deduced from a test based on household budget allocation decisions akin to a test of separability in a traditional demand system setting (Allais et al 2015).

8.4 Estimation Methods

When the choice probabilities have a closed form expression, we can easily use the maximum likelihood method to estimate the parameters θ. Define $P_{ht}(\theta)$ as the probability that the consumer h chooses any alternative or a bundle of alternatives on purchase occasion t. The probability of the sequence of observed choices of consumer h is then $S_h(\theta) = \prod_{t=1}^{T} P_{ht}(\theta)$ and assuming that each consumer's choice is independent of that of other consumers, the log likelihood function could be written as $LL(\theta) = \sum_{h=1}^{H} \ln S_h(\theta)$.

When unobserved heterogeneity in consumer preferences are introduced via random parameters, the choice probabilities no longer have a closed-form expression. The log likelihood function is then a multiple integral that we cannot be solved analytically. In this case, SML is necessary (Train 2003). SML approximates choice probabilities for any given value of θ using the following algorithm: First, we take R random draws from the chosen distributions and compute the simulated probability $SP_{ht}(\theta) = \frac{1}{R}\sum_{r=1}^{R} P_{ht}(\theta^r)$. Second, the simulated likelihood function is then calculated as: $SLL(\theta) = \sum_{h=1}^{H} \ln\left(\prod_{t=1}^{T} SP_{ht}(\theta)\right)$ and can be optimized in a third step. If R rises faster than $\sqrt{HT}$, the maximum likelihood estimator is

consistent, asymptotically normal and efficient, and equivalent to maximum likelihood. In practice, a large number of random draws are needed, and a large number of simulations is typically very computationally expensive. To reduce the number of simulations, randomized and scrambled Halton sequences are often used (Bhat, 2003), where the simulation error falls with the number of Halton draws.

Random utility models such as those presented in this chapter are consistently estimated if the observed characteristics of alternatives b_j are independent from the error term ε_{hj} in each baseline utility function. If we assume $\varepsilon_{hj} = \xi_j + e_{hj}$,where ξ_j is the unobserved term that captures all unobserved product characteristics and e_{hjt} is an individual-specific error term, the independence assumption cannot hold if unobserved factors included in ξ_j (and then included in the error term ε_{hj}) are correlated with observed factors (included in b_j). In this case, the estimated impact of the observed factor captures not only that factor's effect, but also the effect of the correlated, unobserved factor. Unobserved product characteristics could include attributes that are not measured, or marketing efforts such as advertising, sales promotions, shelf position that are observed by the retailer, but not the econometrician. The resulting endogeneity means that all parameter estimates will be biased and inconsistent. For example, if the unobserved factor is advertising, we know that firms maximize profits with respect to both price and advertising so, in general, these decisions cannot be independent. Firms might raise the price of their products when they advertise if they believe that doing so stimulates demand. Alternatively, firms may lower price when they advertise (e.g., as a part of a sale), so the possibility of either case makes the sign of the bias ambiguous.

Endogeneity in discrete choice models is typically addressed through the control function approach (Petrin and Train 2010). Define the vector of observed product attributes as: $b_j = (b_j^0 , y_{jh})$ where b_j^0 is the vector of exogenous product attributes and y_{jh} the endogenous variable. The control function method is a two-step approach in which the endogenous variable y_{jh} is regressed on the exogenous product attributes x_{jht} and instrumental variables Z_j in the first-stage. If the first-stage model is written as: $y_j = Z_j\gamma + b_j^0 \tau + \varpi_j$, then ϖ_{jh} is

the error term. Assuming a joint normal distribution between ϖ and ξ_j, we can re-write the indirect utility function as: $U_{jh} = V(b_j\ , \theta) + \lambda\widehat{\varpi}_j + \sigma\eta_j + \vartheta_{jh}$ where η_j is a standard normal distributed variable and σ is the associated standard deviation. The estimated error term $\widehat{\varpi}_j$ includes some omitted variables that are correlated with the endogeneous variable y_j and not captured by the other exogenous variables of the demand equation b_j^0 or by the instrumental variables Z_j. Introducing this term in the indirect utility function captures unobserved product characteristics that vary across time, and essentially purge the equation of bias as the endogenous variable y_j is now uncorrelated with the new error term $\vartheta_{jh} = \xi_j + e_{hj} - \lambda\widehat{\varpi}_j$. However, because the demand model contains variables that are themselves estimated, the standard errors of the estimated demand parameters must be adjusted accordingly (Karaca-Mandic and Train 2003).

The choice of instrumental variables Z_j is crucial. Good instruments must be independent of the error term ξ_j, make economic sense, be sufficiently correlated with the endogenous regressors, but must not be correlated between themselves. In order to control for price endogeneity, three kinds of instruments are generally used. First, input prices are generally uncorrelated with customer choices, but are correlated with prices from the theory of the firm (Bonnet and Dubois 2010). Assuming no spatial correlation between markets, prices in other markets can also be valid proxies for the cost of production (Hausman, Leonard, and Zona 1994; Nevo 2000). Finally, attributes of other products are not correlated with the demand for the product in question, but are likely to be correlated with its price (Berry et al. 1995). If other variables are thought to be endogenous, then similar instruments must be found. For example, Richards and Hamilton (2015) instrument for endogenous variety, while Allais et al. (2015) instrument for label choices.

9 Conclusions and Implications

In this chapter, we review a broad selection of methods that have been used to study problems in consumer demand over the last 20 years, and provide a hint as to the types of models

likely to be used in the near future. In each case, the form of the model is driven by both the type of data that are available, and the question at hand. While most practical applications of these models involve demand elasticities, they are equally adept at producing demand forecasts, or for inference and drawing conclusions regarding the causal effect of a policy treatment. As computing power and data gathering capabilities advance, our methods will surely keep pace.

References

[1] Ainslie, A., and P. E. Rossi. 1998. Similarities in choice behavior across product categories. Marketing Science 17: 91-106.

[2] Allais O., F. Etilé and S. Lecocq 2015. Mandatory labels, taxes and market forces: An empirical evaluationof fat policies. Journal of Health Economics. 43: 27-44

[3] Anderson, S. P., A. de Palma, and J. F. Thisse. 1992. Discrete choice theory of product differentiation. MIT press.

[4] Anselin, L. 2002. Under the hood issues in the specification and interpretation of spatial regression models. Agricultural Economics 27(3): 247-267.

[5] Athey, S., and G. W. Imbens. 2015. Machine learning methods for estimating heterogeneous causal effects. Working paper, Graduate School of Business, Stanford University, Stanford, CA.

[6] Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. 2015. Machine learning methods for demand estimation. American Economic Review. 105: 481-85.

[7] Bell, D. R., and J. M. Lattin. 1998. Shopping behavior and consumer preference for store price format: Why "large basket" shoppers prefer EDLP. Marketing Science 17: 66-88.

[8] Berry, S. T., J. Levinsohn, and A. Pakes. 1995. Automobile prices in market equilibrium. Econometrica 63(4): 841-890.

[9] Berry, S. T. 1994. Estimating discrete-choice models of product differentiation. The RAND Journal of Economics 25(2): 242-262.

[10] Besanko, D., S. Gupta, and D. Jain. 1998. Logit demand estimation under competitive pricing behavior: An equilibrium framework. Management Science 44(11): 1533-1547.

[11] Bell, D. R., T. H. Ho, and C. S. Tang. 1998. Determining where to shop: Fixed and variable costs of shopping. Journal of Marketing Research 35: 352-369.

[12] Bhat, C. R. 2000. Incorporating observed and unobserved heterogeneity in urban work mode choice modeling Transportation Science, 34:228–238.

[13] Bhat, C. R. 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. Transportation Research Part B: Methodological 37(9): 837-855.

[14] Bhat, C. R. 2005. A multiple discrete–continuous extreme value model: formulation and application to discretionary time-use decisions. Transportation Research Part B: Methodological 39(8): 679-707.

[15] Bhat, C. R. 2008. The multiple discrete-continuous extreme value (MDCEV) model: role of utility function parameters, identification considerations, and model extensions. Transportation Research Part B: Methodological 42(3): 274-303.

[16] Bhat, C. R., M. Castro, and A. R. Pinjari. 2015. Allowing for complementarity and rich substitution patterns in multiple discrete–continuous models. Transportation Research Part B: Methodological 81: 59-77.

[17] Belloni, A., V. Chernozhukov, and C. Hansen. 2014. Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies. 81: 608-650.

[18] Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society. Series B (Methodological), 36: 192-236.

[19] Berto Villas-Boas S. 2007.Vertical relationships between manufacturers and retailers : Inference with limited data. Review of Economic Studies, 74:625–652.

[20] Bonnet C. and V. Réquillart. 2013. Tax incidence with strategic firms on the soft drink market, Journal of Public Economics, 106: 77-88

[21] Bonnet C. and P. Dubois 2010. Inference on Vertical contracts between manufacturers and retailers allowing for non-linear pricing and resale price maintenance. RAND Journal of Economics. 41(1): 139-164

[22] Breiman, L. 1996. Bagging predictors. Machine learning, 24(2), 123-140.

[23] Brenkers, R., and F. Verboven. 2006 Market definition with differentiated products: lessons from the car market. Recent Developments in Antitrust: Theory and Evidence, 153.

[24] Briesch, R. A., P. K. Chintagunta, and E. J. Fox. 2009. How does assortment affect grocery store choice? Journal of Marketing Research 46: 176-189.

[25] Bruno, H. A., and N. J. Vilcassim. 2008. Research note-structural demand estimation with varying product availability Marketing Science 27(6): 1126-1131.

[26] Cardell, N. S. 1997. Variance components structures for the extreme-value and logistic distributions with application to models of heterogeneity. Econometric Theory 13(2): 185-213.

[27] Chintagunta, P. K. 1993. Investigating purchase incidence, brand choice and purchase quantity decisions of households. Marketing Science 12(2): 184-208.

[28] Cox, D. R. 1972. The analysis of multivariate binary data. Journal of the Royal Statistical Society Series C 21: 113–120.

[29] Cressie, N. A.C. 1993. Statistics for spatial data. New York: John Wiley and Sons.

[30] Deaton, A., and J. Muellbauer. 1980. Economics and consumer behavior. Cambridge, UK: Cambridge University Press.

[31] Dubé, J. P. 2004. Multiple discreteness and product differentiation: Demand for carbonated soft drinks. Marketing Science 23(1): 66-81.

[32] Dubin, J. A., and D. L. McFadden. 1984. An econometric analysis of residential electric appliance holdings and consumption. Econometrica 52: 345-362.

[33] Feenstra, R. C., and J. A. Levinsohn. 1995. Estimating markups and market conduct with multidimensional product attributes. Review of Economic Studies 62(1): 19-52.

[34] Foncel J. and M. Ivaldi. 2005. Operating system prices in the home pc market. The Journal of Industrial Economics, LIII(2):265–297

[35] Goeree, M. S. 2008. Limited information and advertising in the US personal computer industry. Econometrica 76(5): 1017-1074.

[36] Goldberg, P. K. 1995. Product differentiation and oligopoly in international markets: The case of the US automobile industry. Econometrica 63: 891-951.

[37] Guadagni, P. M., J. D. Little. 1983. A logit model of brand choice calibrated on scanner data. Marketing Science 2: 203-238.

[38] Hanemann, W. M. 1984. Discrete/continuous models of consumer demand. Econometrica 53: 541-561.

[39] Hendel, I. 1999. Estimating multiple-discrete choice models: An application to computerization returns. Review of Economic Studies 66(2): 423-446.

[40] Hausman, J. A. 1978. Specification tests in econometrics. Econometrica 46: 1251-1271.

[41] Hausman, J., G. Leonard, and J. D. Zona. 1994. Competitive analysis with differenciated products. Annales d'Economie et de Statistique 34: 159-180.

[42] Heckman, J. J. 1979. Sample selection bias as a specification error. Econometrica 48: 153-161.

[43] James, G., D. Witten, T. Hastie, and R. Tibshirani. 2014. An introduction to statistical learning: with applications in R. New York: Springer.

[44] Kamakura, W., and K. Kwak. 2012. Menu-choice modeling. Working paper, Rice University, Department of Marketing.

[45] Karaca-Mandic, P., and K. Train. 2003. Standard error correction in two-stage estimation with nested samples. The Econometrics Journal 6(2): 401-407.

[46] Kelejian, H. H., and I. R. Prucha. 1999. A generalized moments estimator for the autoregressive parameter in a spatial model. International Economic Review 40(2): 509-533.

[47] Kim, J., G. M. Allenby, and P. E. Rossi. 2002. Modeling consumer demand for variety. Marketing Science 21(3): 229-250.

[48] Kwak, K., S. D. Duvvuri, and G. J. Russell. 2015. An analysis of assortment choice in grocery retailing. Journal of Retailing 91: 19-33.

[49] Lancaster, K. J. 1966. A new approach to consumer theory. Journal of Political Economy 74: 132-56.

[50] Lee, L. F., G. S. Maddala, and R. P. Trost. 1980. Asymptotic covariance matrices of two-stage probit and two-stage tobit methods for simultaneous equations models with selectivity. Econometrica 42: 491-503.

[51] Luce, R. D.1959. *Individual Choice Behavior*, New York: John Wiley & Sons, Inc.

[52] Manchanda, P., A. Ansari, S. Gupta. 1999. The shopping basket: A model for multicategory purchase incidence decisions. Marketing Science 18: 95–114.

[53] Mehta, N., S. Rajiv, and K. Srinivasan. 2003. Price uncertainty and consumer search: A structural model of consideration set formation. Marketing Science 22(1): 58-84.

[54] McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior, in *Frontiers in Econometrics*, ed. by P. Zarembka, New York: Academic Press, pp. 105-142

[55] McFadden, D., and K. Train. 2000. Mixed MNL models for discrete response. Journal of Applied Econometrics 15(5): 447-470.

[56] Moon, S. and G. J. Russell. 2008. Predicting product purchase from inferred customer similarity: an autologistic model approach. Management Science 54: 71–82.

[57] Nevo, A. 2001. Measuring market power in the ready-to-eat cereal industry. Econometrica 69(2): 307-342.

[58] Petrin, A., and K. Train. 2010. A control function approach to endogeneity in consumer choice models. Journal of Marketing Research 47(1): 3-13.

[59] Phaneuf, D. J., C. L. Kling, and J. A. Herriges. 2000. Estimation and welfare calculations in a generalized corner solution model with an application to recreation demand. Review of Economics and Statistics 82(1): 83-92.

[60] Pinkse, J., M. E. Slade, and C. Brett. 2002. Spatial price competition: a semiparametric approach. Econometrica 70: 1111-1153.

[61] Pinkse, J., and M. E. Slade. 2004. Mergers, brand competition, and the price of a pint. European Economic Review 48(3): 617-643.

[62] Pofahl, G. M., and T. J. Richards. 2009. Valuation of new products in attribute space. American Journal of Agricultural Economics 91(2): 402-415.

[63] Richards, T. J. 2000. A discrete/continuous model of fruit promotion, advertising, and response segmentation. Agribusiness 16(2): 179-196.

[64] Richards, T. J., M. I. Gómez, and G. F. Pofahl. 2012. A multiple-discrete/continuous model of price promotion Journal of Retailing 88(2): 206-225.

[65] Richards, T. J., S. F. Hamilton, and W. J. Allender. 2014. Social networks and new product choice. American Journal of Agricultural Economics 96(2): 489-516.

[66] Richards, T. J. and S. F. Hamilton. 2015. Variety pass-through: An examination of the ready-to-eat cereal market. forthcoming in the Review of Economics and Statistics.

[67] Rojas, C., and E. B. Peterson. 2008. Demand for differentiated products: Price and advertising evidence from the US beer market. International Journal of Industrial Organization 26(1): 288-307.

[68] Russell, G. J. and A. Petersen. 2000. Analysis of cross category dependence in market basket selection. Journal of Retailing 76: 367–92.

[69] Slade, M. E. 2004. Market power and joint dominance in UK brewing. The Journal of Industrial Economics 52: 133-163.

[70] Smith, H. 2004. Supermarket choice and supermarket competition in market equilibrium. The Review of Economic Studies 71: 235-263.

[71] Song, I. and P. K. Chintagunta. 2006. Measuring cross-category price effects with aggregate store data. Management Science 52: 1594–609.

[72] Train, K. E. 2003. Discrete choice methods with simulation. Cambridge University Press.

[73] Varian, H. R. 2014. Big data: New tricks for econometrics. Journal of Economic Perspectives. 3-27.

[74] Vásquez-Lavín, F., and M. Hanemann. 2008. Functional forms in discrete / continuous choice models with general corner solution. Department of Agricultural and Resource Economics, University of California Berkeley, CUDARE Working Paper 1078.

[75] Villas-Boas, S. B. 2007. Vertical relationships between manufacturers and retailers: Inference with limited data. Review of Economic Studies 74: 625-652.

[76] Wales, T. J., and A. D. Woodland. 1983. Estimation of consumer demand systems with binding non-negativity constraints. Journal of Econometrics 21(3): 263-285.