Panel Data Discrete Choice Models of Consumer Demand

By

Michael P. Keane

University of Oxford, Department of Economics and Nuffield College

Prepared for *The Oxford Handbooks*: *Panel Data*

June 3, 2013

Acknowledgements: Keane's work on this project was supported by Australian Research Council grants FF0561843 and FL110100247.

1. Introduction

This chapter deals with the vast literature on panel data discrete choice models of consumer demand. The reason this area is so active is that very high quality data is available. Firms like Nielsen and IRI have, for over 30 years, been collecting panel data on households' purchases of consumer goods. This is known as "scanner data," because it is collected by check-out machine scanners. Available scanner data sets often follow households for several years, and record all their purchases in several different product categories. The typical data set not only contains information on the universal product codes (UPC) of the consumer goods that households buy on each shopping trip, but also information on several exogenous forcing variables, such as price and whether the goods were displayed or advertised in various ways.

To my knowledge the first paper using scanner data to study the impact of price and other marketing variables on consumer demand was Guadagni and Little (1983) in *Marketing Science*. But few economists knew about scanner data until the mid to late 90s. Once they became aware of this treasure trove of data, they started to use it very actively. Today, estimation of demand models on scanner data has become a major part of the field of empirical industrial organization.

Thus, the consumer demand literature based on scanner data is unusual relative to other literatures discussed in this *Handbook* in two respects. First, it remains true that the majority of work in this area is by marketers rather than economists. Second, this is an uncommon case where the "imperial science" of economics (see, e.g., Stigler (1984)) has experienced a substantial knowledge transfer from another area (i.e., marketing). Furthermore, it should be noted that discrete choice models of consumer demand are also widely used in other fields like transportation research, agricultural and resource economics, environmental economics, etc..

Given that the literature on panel data models of consumer demand is so large, I will make no attempt to survey all the important papers in the field. Instead, I will focus on the main research questions that dominate this area, and the progress that has been made in addressing them. Thus, I apologize in advance for the many important papers that are not cited.

The most salient feature of scanner panel data is that consumers exhibit substantial persistence in their brand choices. In the language of marketing, consumers show substantial "brand loyalty." A second obvious aspect of the data is that, if we aggregate to the store level, then in most product categories the sales of a brand jump considerably when it is on sale (i.e., typically the price elasticity of demand is on the order of 3 to 5). Superficially these two

observations *seem* contradictory. If individual consumers are very loyal to particular brands, then why would demand for brands be very price sensitive in the aggregate?

In light of these empirical observations, the first main objective of the panel data demand literature has been to understand the underlying sources of persistence in brand choices. Based on work by Heckman (1981) on employment dynamics, it is now understood that persistence in brand choices may arise from three sources: (i) permanent unobserved heterogeneity in tastes, (ii) serial correlation in taste shocks, or (iii) "true" or "structural" state dependence.

Only the third source of persistence (i.e., state dependence) involves a causal effect of past choices on the current choice (and, likewise, an effect of the current choice on future choices). Uncovering whether state dependence exists is of great importance in both marketing and industrial organization. If it exists, then current marketing actions, such as price discounts, will affect not only current but also future demand. This has important implications for pricing policy, the nature of firm competition, etc..

The second major objective of the literature has been to distinguish alternative possible explanations for structural state dependence (assuming that it exists). Some of the potential explanations include habit persistence, learning about quality through trial, inventory behavior, variety seeking behavior, switching costs, and so on.

A third, but closely related, major objective of the literature has been to understand the dynamics of demand. Most important is to understand the sources of the observed increase in demand when a brand is on sale. The increase in sales may arise from three sources: (i) brand switching, (ii) category expansion, or (iii) purchase acceleration, also known as cannibalization. In everyday language, these correspond to (i) stealing customers from your competitors, (ii) bringing new customers into the category, or (iii) merely accelerating purchases by consumers who are loyal to a brand and who would have eventually bought it at the regular price anyway.

The distinction among these three sources of increased demand is obviously of crucial important for pricing policy. For example, if most of the increase of sales that results from a price discount is due to cannibalization of future sales, then a policy of having periodic price discounts obviously makes no sense.

The estimation of discrete choice models with many alternatives is a difficult econometric problem. This is because the order of integration required to calculate choice probabilities in such a model is typically on the order of J-1, where J is the number of choice

alternatives. The development of simulation methods for the estimation of multinomial discrete choice models in the late 80s was largely motivated by this problem (see McFadden (1989)).

As discussed in Keane (1994), in the panel data case the required order of integration to construct the choice probabilities in discrete choice models is much higher. This is because it is the probability of a consumer's entire choice sequence that enters the likelihood function. Thus, the required order of integration is (J-1)·T, where T is the number of time periods. In typical scanner panels T is on the order of 50 to 200 weeks, so the order of integration is very high.

In Keane (1994), I developed a method of "sequential importance sampling" that makes estimation of panel data discrete choice models feasible. In the special case of the normal errors, which gives the panel probit model, this method is known as the "GHK" algorithm. GHK is a highly accurate method for approximating multi-dimensional normal integrals. It is notable that the development of simulation based econometric methods has gone hand-in-hand with the desire to estimate demand models with large choice sets, multiple time periods, and complex error structures.

The outline of the remainder of the chapter is as follows. In section 2, I describe a fairly general panel data discrete choice model. Section 3 discusses the econometric methods needed to estimate such models. Then, Section 4 discusses the theoretical issues involved in distinguishing state dependence from heterogeneity, while Section 5 discusses empirical work on state dependence and/or choice dynamics. Section 6 concludes.

2. The Typical Structure of Panel Data Discrete Choice Models

Here I describe the typical structure of demand models used in marketing (and more recently in industrial organization). Let j =1,…,J index alternatives, t =1,…,T index time, and i =1,…,N index people. Then the "canonical" brand choice model can be written as follows:

$$U_{ijt} = \alpha_{ij} + X_{ijt}\beta + \gamma d_{ij,t-1} + \varepsilon_{ijt} \quad where \quad \varepsilon_{ijt} = \rho\varepsilon_{ij,t-1} + \eta_{ijt} \tag{1}$$

$$d_{ijt} = 1 \;\; if\; U_{ijt} > U_{ikt}\; for\; all\; k \neq j \;\; d_{ijt} = 0 \;\; otherwise \tag{2}$$

Equation (1) expresses the utility that person i receives from the purchase of brand j at time t. Utility (U_{ijt}) depends on a vector of product attributes X_{ijt} and the utility or attribute weights β. Utility also depends on consumer i's intrinsic preference for brand j, which I denote by α_{ij}. It is further assumed that utility depends on whether brand j was chosen by person i on the previous

choice occasion ($d_{ij,t-1}$=1). Finally, there is a purely idiosyncratic person, time and brand specific taste shock ε_{ijt}. This is allowed to be serially correlated, with the fundamental shocks η_{ijt} being *iid*. Equation (2) simply says that person *i* chooses the brand *j* that gives him greatest utility at time *t*. Of course, in a discrete choice model we only observe choices and not utilities.

Before turning to the econometrics it is important to give an economic interpretation to the terms in (1). A utility function that is linear in attributes is quite standard in the demand literature (see Lancaster (1966) for the classic exposition of attribute based utility). But in (1) we assume the utility weights β are common across consumers (as in traditional logit and probit models). This is a strong assumption, but it is only for expositional convenience.[1] The simulation methods discussed below can easily accommodate heterogeneity in β.

I will focus attention on heterogeneity in the brand intercepts α_{ij}. These capture consumer heterogeneity in tastes for attributes of alternatives that are not observed by the econometrician (see Berry (1994), Elrod and Keane (1995), Keane (1997)). For example, in some products like cars or clothing or perfume, different brands convey a certain "image" that is hard to quantify. Heterogeneous tastes for that "image" would be subsumed in the α_{ij}. Of course, even mundane products have unobserved attributes (e.g., the "crispness" of different potato chips).

It is worth emphasizing that one of the attributes included in X_{ijt} is price, which we denote by p_{ijt}. The budget constraint conditional on purchase of brand *j* is $C_{it} = I_{it} - p_{ijt}$. As frequently purchased consumer goods are fairly inexpensive, it makes sense to assume the marginal utility of consumption of the outside good is a constant over the range $[I_{it}-p_{max}, I_{it}-p_{min}]$, where p_{max} and p_{min} are the highest and lowest prices ever observed in the category. This justifies making utility linear in consumption of the outside good. If we use the budget constraint to substitute for C_{it}, we obtain a conditional indirect utility function that is linear in income and price.

Furthermore, income is person specific and not alternative specific. Because income is the same across all alternatives *j* for an individual, it does not alter the utility differences between alternatives. As a result, income drops out of the model and we are left with only price. It is important to remember, however, that price only appears because we are dealing with an indirect utility function, and its coefficient is not interpretable as just another attribute weight. The price coefficient is actually the marginal utility of consumption of the outside good.

[1] Product attributes can be "vertical" or "horizontal." A vertical attribute is something like quality that all consumers would like more off. A horizontal attribute is something like saltiness of crackers, which some people would like and others dislike. Thus, for horizontal attributes, even the sign of β may differ across consumers.

Thus, an important implication of consumer theory is that the price coefficient should be equal across all alternatives. However, it will generally vary across people, as the marginal utility of consumption is smaller for those with higher income. This can be accounted for by letting the price coefficient depend on income and other household characteristics.

The next important feature of (1) is the lagged choice variable $d_{ij,t-1}$. This captures an effect of lagged purchase of a brand on its current utility evaluation. Heckman (1981) calls this "structural" state dependence. Most papers use more elaborate forms of state dependence than just lagged purchase. For instance, Guadagni and Little (1983) used an exponentially smoothed weighted average of all the lagged d_{ijs} for s=1,…,t-1, and this specification is popular in the marketing literature. But I will focus on the first-order Markov model for expositional purposes.

There are many reasons why a structural effect of lagged purchase on current utility may exist; such as habit persistence, learning, inventories, variety seeking behavior, switching costs and so on. I discuss efforts to distinguish among these sources of state dependence in Section 5.

First, in Section 4, I'll focus on the question of whether state dependence exists at all (whether $\gamma \neq 0$). This question alone has been the focus of a large literature. The question is difficult to address, because failure to adequately control for heterogeneity and other serial correlation will lead to what Heckman (1981) called "spurious" state dependence. Furthermore, there are deep econometric and philosophical issues around the question of whether it is even possible to distinguish state dependence from heterogeneity (or serial correlation in general).

Finally, equation (1) includes idiosyncratic taste shocks ε_{ijt}. These may be interpreted in different ways, depending on ones perspective. In the economic theory of random utility models (Bloch and Marschak (1960), McFadden (1974)) choice is deterministic from the point of view of the consumer, who observes his/her own utility. In that case, choice only appears to be random from the point of view of the econometrician, who has incomplete information about consumer preferences and brand attributes. As Keane (1997) discusses, the ε_{ijt} can be interpreted as arising from unobserved attributes of brands for which people have heterogeneous tastes that vary over time. This is in contrast to the brand intercepts α_{ij}, which capture unobserved attributes of brands for which people have heterogeneous tastes that are constant over time. However, in psychology- based models of choice, the ε_{ijt} are interpreted as genuinely random elements of choice behavior. I am not aware of a convincing way to distinguish between these two perspectives.

If the ε_{ijt} arise from time-varying tastes, it is plausible that tastes show some persistence over time. This motivates the AR(1) specification $\varepsilon_{ijt} = \rho\varepsilon_{ij,t-1} + \eta_{ijt}$ where η_{ijt} is *iid* over time and people. If ρ>0 then taste shocks exhibit temporal persistence.

If the η_{ijt} are correlated across brands it implies some brands are more similar than others on the unobserved attribute dimensions for which people have time-varying tastes. Similarly, if the intercepts α_{ij} are correlated across brands it implies some brands are more similar than others on the unobserved attribute dimensions for which people have time-invariant tastes. Brands that are more similar on the latent attribute dimensions, will *ceteris paribus*, have more switching between them and higher cross-price elasticities of demand.

These ideas are the basis of the "market mapping" literature that uses panel data to determine the location of brands in a latent attribute space (see Elrod (1988), Elrod and Keane (1995), Keane (1997)). For example, in a market map for cars, Mercedes and BMW would presumably lie close together in one part of the space, while Ford and Chevy trucks would also lie close together but in a very different part of the space. An estimated market map can, for example, help a firm to determine who its closest competitors are.

Note that the multinomial logit model assumes all errors are uncorrelated. This makes all brands "equally (dis)similar" (i.e., equally spread out in the market map) so that all cross-price elasticities of demand are equal. It was a desire to escape this unrealistic assumption that resulted in work on simulation methods – see, e.g., Lerman and Manski (1981) and McFadden (1989) – that make estimation of the multinomial probit model (with correlated normal errors) feasible. This is the focus of the next section.

3. Estimation of Panel Data Discrete Choice Models

Here I discuss the computational problems that arise in estimating panel data discrete choice models. Maximum likelihood estimation of the model in (1)-(2) requires distributional assumptions on the intercepts α_{ij} and the errors η_{ijt}. The most common assumptions in the literature are that the intercepts are either multivariate normal ($\alpha_i \sim N(0,\Sigma)$) or multinomial, while the η_{ijt} are either normal ($\eta_{it} \sim N(0,\Omega)$) or *iid* type I extreme value. If both the α_{ij} and η_{ijt} are normal we have the random effects panel probit model. If the α_{ij} are normal while the η_{ijt} are extreme value we have a normal mixture of logits model (N-MIXL). If the α_{ij} are multinomial we have a discrete mixture of probits or logits. These are often called "latent class" models.

Estimation of the model in (1)-(2) requires some identifying normalizations. In discrete choice models, there is no natural scale for utility, and only utility differences among alternatives determine choices. Thus, one alternative (often but not always a "no purchase" option) is chosen as the base alternative, and its utility is normalized to zero. Hence, the error covariance matrices Σ and Ω are of rank (J-1) rather than J. The scale of utility is usually fixed by letting the idiosyncratic errors η_{ijt} be *standard* normal or *standard* type I extreme value.

Now, consider the panel probit case. In order to form the likelihood for a person *i* we need to form the probability of his/her observed sequence of choices given the observed vector of covariates. That is, we need $P(d_{ij(1),1}, \ldots, d_{ij(T),T}|X_{i1}, \ldots, X_{iT})$, where $j(t)$ denotes the index j of the option that the consumer actually chose at time *t*, while the $X_{it} \equiv (x_{i1t},\ldots,x_{iJt})$ are vectors of covariates for all *J* alternatives at time *t*. The difficulty here is that, given the structure (1)-(2), this joint probability is very computationally difficult to construct.

First, consider the case where $\gamma=\rho=0$. That is, there is no state dependence and the idiosyncratic errors ε_{ijt} are serially independent. Then the only source of persistence in choices over time are the brand specific individual effects ($\alpha_{i1},\ldots, \alpha_{iJ}$). This gives an equicorrelated structure for the composite error terms $v_{ijt} = \alpha_{ij} + \varepsilon_{ijt}$, so we have a "random effects probit model." Here, choice probabilities are independent over time *conditional* on the α_{ij}, so we have:

$$P\left(d_{ij(1),1}, \ldots, d_{ij(T),T}|X_{i1}, \ldots, X_{iT}\right) = \int_{-\infty}^{\infty} \prod_{t=1}^{T} P\left(d_{ij(t),t}|X_{it}, \alpha\right) f(\alpha|\Sigma) d\alpha \qquad (3)$$

Each conditional probability $P\left(d_{ij(t),t}|X_{it}, \alpha_i\right)$ is a cross-section probit probability. As is well known, these are multivariate normal integrals of dimension *J*-1. When $J \geq 3$ or 4, it is necessary to use simulation methods like the GHK algorithm to evaluate these integrals. As the focus here is on panel data issues and not problems that already arise in cross-section discrete choice models, I'll refer the reader to Geweke and Keane (2001) for further details.

The key problem in forming the choice probability in (3) is how to evaluate the integral over the density $f(\alpha|\Sigma)$ of the multivariate normal (*J*-1)-vector of individual effects α. Butler and Moffitt (19**) proposed a computationally efficient Gaussian quadrature procedure to evaluate normal integrals like that in (3). The procedure involves replacing the integral in (3) with a weighted sum over Gauss-Hermite quadrature points. If J=2, so α is a scalar, we have:

$$\hat{P}_{Q,G}\left(d_{ij(1),1}, \ldots, d_{ij(T),T}|X_{i1}, \ldots, X_{iT}\right) = \sum_{g=1}^{G} \prod_{t=1}^{T} w_g P\left(d_{ij(t),t1}|X_{it}, \alpha_g\right) \qquad (4)$$

The α_g are the quadrature points, and the w_g are the associated weights. Butler and Moffitt (1982) describe the derivation of the weights and points, and find that rather accurate evaluations of normal integrals can be obtained using just several points (i.e., typically only 6 or 7).
In the case of J=3 one needs two sets of quadrature points (α_{i1},α_{i2}) and the single sum in (4) is replaced by a double sum. In general, a J-1 dimensional sum is required. Thus, quadrature, like other numerical methods for evaluating integrals, suffers a curse of dimensionality. As a result, the quadrature method is applicable when J is fairly small (i.e., J≤3).

However, a useful strategy when J is large is to impose a relatively low dimensional factor structure on Σ. Then the required order of integration in (3) is the number of factors (F) regardless of the size of J-1. Lancaster (1963) discussed the idea that in a market with many products, those products may only be differentiated on a few attribute dimensions (e.g., there are hundreds of brands of cereal, but they differ on only a few attributes like sugar content, fibre content, etc.). Work on "market mapping" using scanner data finds that the unobserved attribute space for most products is well described by just a few factors (e.g., F≤3), even when J is very large (see Elrod (1984), Elrod and Keane (1995), Keane (1997), Andrews and Manrai (1999)).

Another advantage of using a factor structure with F<J-1 is that, in general, the number of parameters in Σ is J·(J-1)/2. Even for modest J it is cumbersome to estimate so many parameters. And, although formally identified, estimation of large covariance matrices creates severe practical/numerical problems in discrete choice models (see Keane (1992), Keane and Wasi (2013)). But in a model with F factors the number of factor loadings to be estimated is F·(J-1), which increases only linearly with J, thus breaking the curse of dimensionality.

Given the speed of modern computers, a brute force frequency simulation approach is also feasible, even when J is very large. That is, let $\{\alpha_d\}_{d=1,\ldots,D}$ denote D draws from the $f(\alpha|\Sigma)$ density obtained using a random number generator. This gives:

$$\hat{P}_{F,D}\left(d_{ij(1),1},\ldots,d_{ij(T),T}|X_{i1},\ldots,X_{iT}\right)=\frac{1}{D}\sum_{d=1}^{D}\prod_{t=1}^{T}P\left(d_{ij(t),t}|X_{it},\alpha_d\right) \quad (5)$$

The similarity between (4) and (5) is notable, as each involves evaluating the choice probabilities at a discrete set of α values and summing the results. The difference is that the quadrature points are chosen analytically so as to provide an accurate approximation with as few points as possible, while in (5) the α_d are simply drawn at random. This means that the number of draws D must be quite large to achieve reasonable accuracy (i.e., at least a few hundred in most applications).

However, the virtue of simulation is that, unlike quadrature and other numerical methods, it does not suffer from the curse of dimensionality. The simulation error variance in simulation estimators of probabilities is of order 1/D, regardless of the size of J. That is, in equation (5) we have $\hat{P}_{F,D} - P \backsim N(0, s^2/D)$ by the Central Limit Theorem, where $s^2 = E\left(\hat{P}_{F,1} - P\right)^2$ and it is assumed that the simulation errors are *iid* across draws *d*. Simulation also substitutes machine time for human time, as in complex models with J>2 the quadrature points can be cumbersome to derive analytically.

In any simulation estimation method it is important that one hold draws fixed as one iterates on the model parameters. Failure to do so creates two related problems: (i) the simulated likelihood will "jump" when the draws change, so the change in the likelihood is not solely due to updating of parameters, (ii) such draw induced changes in the simulated likelihood play havoc with the calculation of likelihood derivatives and the operation of parameter search algorithms. But holding the draws $\{\alpha_d\}_{d=1,\ldots,D}$ fixed would appear to be impossible in the random effects probit model, because as Σ changes it seems one must take new draws for α from the new $f(\alpha|\Sigma)$. A standard "trick" that can be used to hold draws fixed as Σ changes works as follows: First let Σ = AA', where A is the lower triangular Cholesky matrix. Then let $\alpha = A\mu$ where μ is a standard normal vector. The "trick" is to draw μ rather than α, and hold the draws $\{\mu_d\}_{d=1,\ldots,D}$ fixed as we iterate on the elements of A. Then the draws $\{\alpha_d\}_{d=1,\ldots,D}$ will vary smoothly as we vary A, causing $\hat{P}_{F,D}$ to vary smoothly. This procedure has the added benefit that iteration on elements of A rather than the elements of Σ guarantees that $\hat{\Sigma}$ will always be a positive definite covariance matrix (by definition of the Cholesky transform).

A more sophisticated way to simulate the integral in (3) is to use sequential importance sampling, as developed in Keane (1993, 1994). This approach, known as the "GHK" algorithm in the special case of importance sampling from the normal, is described in quite a few papers in the literature,[2] so I just give a basic example here. Continue to consider the case of γ=ρ=0, and define the composite error:

$$v_{ijt} = \alpha_{ij} + \varepsilon_{ijt} \tag{6}$$

Equation (1) implies a bound on $v_{ijt} = \alpha_{ij} + \varepsilon_{ijt}$ such that option *j* is chosen at time *t*:

[2] Aside from the two papers cited in the text, see also Hajivassiliou, McFadden and Ruud (1996), Geweke, Keane and Runkle (1994, 1997), Geweke and Keane (2001).

$$U_{ijt} > U_{ikt}\ \forall\ k \neq j \quad \Rightarrow \quad \nu_{ijt} \geq -X_{ijt}\beta + (X_{ikt}\beta + \nu_{ikt})\ \forall\ k \neq j \tag{7}$$

To simplify even further, consider the case where J=2. As we noted earlier, the utility of a base option (say #1) is normalized to zero, leaving a single utility index U_{it} for the other option (say #2). Hence we do not need the *j* subscript in this case. We write that *j*=2 is chosen over *j*=1 *iff*:

$$U_{it} > 0 \quad \Rightarrow \quad \nu_{it} \geq -X_{it}\beta \tag{7'}$$

Now, to be concrete, consider the problem of simulating the probability of a particular sequence $(d_{i1} = 2, d_{i2} = 2, d_{i3} = 2)$. That is, T=3 and the consumer chooses option 2 in all three periods.

To implement the GHK algorithm we divide the sequence probability into transition probabilities. That is, we have:

$$P(d_{i1} = 2, d_{i2} = 2, d_{i3} = 2|X_{i1}, \dots, X_{i3}) = P(d_{i1} = 2|X_{i1})$$
$$P(d_{i2} = 2|d_{i1} = 2, X_{i1}, X_{i2})P(d_{i3} = 2|d_{i1} = 2, d_{i2} = 2, X_{i1}, X_{i2}, X_{i3}) \tag{8}$$

A key point is that the transition probabilities in (8) depend on lagged choices and covariates despite the fact that we have assumed γ=0, so there is no true state dependence (only serial correlation). This occurs because of a fundamental property of discrete choice models:

Specifically, as we only observe choices and not the latent utilities, we cannot construct lagged values of the error term. For instance, if $d_{i1} = 2$, all this tells us is that $\nu_{i1} \geq -X_{i1}\beta$. Thus we cannot form the transition probability $P(d_{i2} = 2|\nu_{i1}, X_{i2})$. We can only form:

$$P(d_{i2} = 2|d_{i1} = 2, X_{i1}, X_{i2}) = P(d_{i2} = 2|\nu_{i1} \geq -X_{i1}\beta, X_{i2}) \tag{9}$$

Notice that both the lagged choice and lagged covariates are informative about the distribution of ν_{i2} as they enable us to infer its truncation (i.e., $\nu_{i1} \geq -X_{i1}\beta$). And, given that the errors are serially correlated, we have a conditional density of the form $f(\nu_{i2}\,|\nu_{i1} \geq -X_{i1}\beta)$.

The computational problem that arises in discrete choice panel data models becomes obvious when we move to period 3. Now, the fact that $(d_{i1} = 2, d_{i2} = 2)$ only tells us that $\nu_{i1} \geq -X_{i1}\beta$ and $\nu_{i2} \geq -X_{i2}\beta$. We have that:

$$P(d_{i3} = 2|d_{i1} = 2, d_{i2} = 2, X_{i1}, X_{i2}, X_{i3}) = P(d_{i3} = 2|\nu_{i1} \geq -X_{i1}\beta, \nu_{i2} \geq -X_{i2}\beta, X_{i3}) \tag{10}$$

The point is that the history at *t*=1 still matters for the *t*=3 choice probability, because of the fact that $\nu_{i1} \geq -X_{i1}\beta_1$ contains additional information about the distribution of ν_{i3} beyond that contained in the *t*=2 outcome, $\nu_{i2} \geq -X_{i2}\beta$. Thus, the conditional density of ν_{i3} has the form $f(\nu_{i3} \,|\nu_{i1} \geq -X_{i1}\beta, \nu_{i2} \geq -X_{i2}\beta)$. And the probability of the sequence (2, 2, 2) is:

$$\int_{-X_{i1}\beta}^{\infty}\int_{-X_{i2}\beta}^{\infty}\int_{-X_{i3}\beta}^{\infty} f(\nu_3 \,|\nu_1 \geq -X_{i1}\beta, \nu_2 \geq -X_{i2}\beta)\, f(\nu_2 \,|\nu_1 \geq -X_{i1}\beta) f(\nu_1) d\nu_3 d\nu_2 d\nu_1 \quad (11)$$

Thus, the probability of a 3 period sequence is a 3-variate integral. And the probability of a T period sequence is a T-variate integral, as the *entire* history matters for the choice probability in any period. If we consider J>2, then the probability of a T period sequence is a T·(J-1) variate integral. This explains the severe computational burden of estimating panel probit models.

This problem is in sharp contrast to a linear model with serially correlated errors, such as:

$$y_{it} = x_{it}\beta + \varepsilon_{it} \quad where \quad \varepsilon_{ijt} = \rho\varepsilon_{ij,t-1} + \eta_{ijt} \quad \eta_{ijt} \sim iid \quad (12)$$

Here we can form E(y_{it} | x_{it}, $\varepsilon_{i,t-1}$) = $x_{it}\beta + \rho\varepsilon_{i,t-1}$ because, conditional on any estimate of β, we observe the lagged error $\varepsilon_{i,t-1} = y_{i,t-1} - x_{i,t-1}\beta$. Similarly, if we could observe ν_{i1} and ν_{i2} in the probit model, then, letting ν_1^* and ν_2^* denote the observed values, equation (11) becomes:

$$\int_{-X_{i1}\beta}^{\infty} f(\nu_1) d\nu_1 \int_{-X_{i2}\beta}^{\infty} f(\nu_2 \,|\nu_1 = \nu_1^*) d\nu_2 \int_{-X_{i1}\beta}^{\infty} f(\nu_3 \,|\nu_1 = \nu_1^*, \nu_2 = \nu_2^*) d\nu_3 \quad (13)$$

Thus the sequence probability would simply be the product of three univariate integrals. The basic idea of the GHK algorithm is to draw values of the unobserved lagged ν_t's and condition on these, enabling us to use equations like (13) to evaluate sequence probabilities rather than (11).

Guided by the structure in (13), the GHK simulator of the sequence probability in (11) is:

$$\hat{P}_{GHK,D}(d_{i1=2}, d_{i2} = 2, d_{i3} = 2|X_i) \quad (14)$$

$$= \frac{1}{D}\sum_{d=1}^{D} \int_{-X_{i1}\beta}^{\infty} f(\nu_1) d\nu_1 \int_{-X_{i2}\beta}^{\infty} f\left(\nu_2 \,\middle|\nu_1 = \nu_1^d\right) d\nu_2 \int_{-X_{i1}\beta}^{\infty} f\left(\nu_3 \,\middle|\nu_1 = \nu_1^d, \nu_2 = \nu_2^d\right) d\nu_3$$

where $\{\nu_1^d, \nu_2^d\}_{d=1}^{D}$ are draws from the *conditional* distributions of ν_1 and ν_2 given that option 2 was chosen in both periods 1 and 2. So GHK replaces the 3-variate integral in (11) by three univariate integrals, and two draws from truncated normal distributions.

A key aspect of the GHK algorithm is how to draw the $\{v_1^d, v_2^d\}_{d=1}^D$sequences in (14) appropriately. The first step is to construct the Cholesky decomposition of the covariance matrix Γ of the error vector (v_{i1}, v_{i2}, v_{i3}). Note that Γ is equicorrelated because the v_{it} have a random effects structure. But the algorithm does not change in any way if Γ has a more complex structure, such as that which would arise if the AR(1) parameter ρ were non-zero. For the Cholesky decomposition we have:

$$\begin{pmatrix} v_{i1} \\ v_{i2} \\ v_{i3} \end{pmatrix} = \begin{pmatrix} 1 & & \\ a_{21} & a_{22} & \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} \eta_{i1} \\ \eta_{i2} \\ \eta_{i3} \end{pmatrix} \tag{15}$$

where a_{11} = 1 to impose that σ_{i1}^2 =1, which is the identifying scale restriction on utility. It is straightforward to draw $\eta_{i1} = v_{i1}$ from a truncated standard normal such that $v_{i1} \geq -X_{i1}\beta$. This can be done by drawing a uniform u_1^d on the interval [$F(-X_{i1}\beta)$, 1] and then setting $\eta_1^d = F^{-1}(u_1^d)$.

Next, we have that $v_{i2} = a_{21}\eta_1^d + a_{22}\eta_2$. Thus, the truncation $v_{i2} \geq -X_{i2}\beta$ implies truncation on η_2 of the form $\eta_{i2} \geq \frac{1}{a_{22}}[-X_{i2}\beta - a_{21}\eta_1^d]$. So we now draw a uniform u_1^d on the interval $F\left(\frac{1}{a_{22}}[-X_{i2}\beta - a_{21}\eta_1^d]\right)$, and set $\eta_2^d = F^{-1}(u_2^d)$. This process can be repeated multiple times for person *i* so as to obtain a set of draw sequences $\{v_{i1}^d, v_{i2}^d\}_{d=1}^D$. Consistent with our earlier discussion, it is the uniform draws $\{u_{i1}^d, u_{i2}^d\}_{d=1}^D$ that should be help fixed as one iterates.

The GHK algorithm can be extended to multiple periods in an obvious way, by adding additional terms to (14). The bound on the time *t* draw is always of the form $\eta_{it} \geq \frac{1}{a_{tt}}[-X_{it}\beta - a_{t1}\eta_1^d - a_{t,t-1}\eta_{t-1}^d]$. With T periods one needs to evaluate T univariate integrals and draw T-1 truncated normals. These operations are extremely fast compared to T-dimensional integration.

Next we consider the case where the α_{ij} are normal while the η_{ijt} are extreme value. This gives the normal mixture of logits model (N-MIXL). It has been studied extensively by Berry (1994), Berry et al (1995), Harris and Keane (1999), McFadden and Train (2000), Train (2003) and others. The choice probabilities have the form:

$$P(d_{ij(t),t}|X_{it}, \alpha_i) = \exp(x_{ij(t),t}\beta + \alpha_{ij(t)}) / \left[1 + \sum_{j=1}^{J} \exp(x_{ijt}\beta + \alpha_{ij})\right] \tag{16}$$

The probability simulator for this model is closely related to the frequency simulator in (5),

except that now we use a logit kernel rather than a probit kernel. As before, let $\{\alpha_d\}_{d=1,\ldots,D}$ denote D random vectors $(\alpha_{1d},\ldots,\alpha_{Jd})$ drawn from the $f(\alpha|\Sigma)$ density, and form the frequency simulator:

$$\hat{P}_{MIXL,D}\left(d_{ij(1),1},\ldots,d_{ij(T),T}|X_{i1},\ldots,X_{iT}\right) = \tag{17}$$

$$\frac{1}{D}\sum_{d=1}^{D}\prod_{t=1}^{T}\exp(x_{ij(t),t}\beta+\alpha_{j(t),d})/\left[1+\sum_{j=1}^{J}\exp(x_{ijt}\beta+\alpha_{jd})\right]$$

One advantage of MIXL is that, in contrast to the random effects probit, once we condition on the individual effects $\alpha_i = (\alpha_{i1},\ldots,\alpha_{iJ})$, the choice probability integrals have a closed form given by the logit kernel $\exp(\cdot)/[1+\exp(\cdot)]$. This makes simulation of the model rather fast and easy.

By introducing correlation across alternatives via the $f(\alpha|\Sigma)$ distribution, the N-MIXL model relaxes the strong IIA assumption of multinomial logit. A number of papers have considered more general distributions for α than the normal. For instance, Geweke and Keane (1999), Rossi, Allenby and McCulloch (2005) and Burda, Harding and Hausman (2008) consider mixture-of-normals models. Indeed, an entire family of MIXL models can be obtained by different choices of the $f(\alpha|\Sigma)$ distribution.

The next set of models that have been popular in the consumer demand literature are "latent class" models. In these models there are a discrete set of consumer types, each with its own vector of brand specific individual effects. That is, we have $(\alpha_{i1},\ldots,\alpha_{iJ}) \in \left(\alpha_1^c,\ldots,\alpha_J^c\right)$ where $c=1,\ldots,C$ indexes types or classes. One estimates both the α^c vector for each class c, as well as the population proportion of each class, π^c. We then obtain unconditional choice sequence probabilities by taking the weight sum over type specific probabilities:

$$P_{LC}\left(d_{ij(1),1},\ldots,d_{ij(T),T}|X_{i1},\ldots,X_{iT}\right) = \sum_{c=1}^{C}\pi_c\prod_{t=1}^{T}P\left(d_{ij(t),t}|X_{it},\alpha^c\right) \tag{18}$$

Here $P\left(d_{ij(t),t}|X_{it},\alpha^c\right)$ is typically a logit or probit kernel. We can interpret the latent class model as a special case of MIXL where the mixing distribution is discrete (in contrast to the normal mixing distributions we considered earlier). Note that the probability in (18) is analytical when a logit kernel is used (no simulation methods are needed).

To my knowledge, Kamakura and Russell (1989) was the first paper to apply the latent class approach in marketing. Work by Elrod and Keane (1995) showed that the latent class approach tends to understate the degree of heterogeneity in consumer preferences. I think it is

fair to say that with the advent of simulation methods, latent class models have become relatively less widely-used (at least in academic research) compared to probit and mixed logit models that allow for continuous heterogeneity distributions.

Recently Keane and Wasi (2013) used several different data sets to compare the fit of latent class models to that of several alternative models with continuous heterogeneity distributions (including N-MIXL and the mixture-of-normals model). We found that models with continuous heterogeneity distributions typically provided a much better fit to the data. Nevertheless, we also found that the simple structure of latent class models often provides useful insights into the structure of heterogeneity in the data, and helps one to understand and interpret results from the more complex models. Thus, it appears that latent class models still have a useful role to play in interpreting discrete choice demand data, even if they are outperformed by other models in terms of fit and predictive ability.

3.B. Extension to Serially Correlated Taste Shocks

So far, I have conducted the discussion of methods for estimating the model in equations (1)-(2) in the case where $\gamma=\rho=0$. That is, there is no state dependence and the idiosyncratic errors (or taste shocks) ε_{ijt} are serially independent. Then the only source of serial correlation was brand specific individual effects. I now consider generalizations of this model. As we discussed in Section 2, it is quite plausible that unobserved brand preferences vary over time rather than being fixed. An example is the AR(1) process in (1). Starting with Keane (1997) and Allenby and Lenk (1994), a number of papers have added AR(1) errors to the random effects structure.

It is simple to discuss extension of the methods we have described to this case of serially correlated ε_{ijt}, as in every case the extension is either simple or practically impossible. For instance, the Butler and Moffitt (1982) quadrature procedure relies specifically on the random effects probit structure and it cannot be extended to serial correlation in ε_{ijt}.

Latent class models are designed specifically to deal with permanent unobserved heterogeneity, so they cannot handle serially correlated idiosyncratic errors. In principle one could have a model with both discrete types and serially correlated idiosyncratic errors. But the resultant models would no longer generate closed form choice probabilities as in (18). They would only be estimable using simulation methods.

On the other hand, the random effects probit model can easily be extended to include serially correlated idiosyncratic errors (like the AR(1) structure in (1)). To estimate this model

using the GHK algorithm, one simply constructs the covariance matrix Γ in a way that incorporates the additional source of serial correlation. Then, construct the corresponding Cholesky matrix and draw the $\{\nu_1^d, \dots, \nu_{T-1}^d\}_{d=1}^D$ sequences in (14) accordingly. Unlike the random effects case, the Γ will no longer be equicorrelated. But the algorithm described in equations (13)-(15) does not change in any way if Γ has a more complex structure.

The frequency simulator in (5) can also be extended to allow for serially correlated ε_{ijt}. For instance, take the model $U_{ijt} = \alpha_{ij} + X_{ijt}\beta + \varepsilon_{ijt}$ where $\varepsilon_{ijt} = \rho\varepsilon_{ij,t-1} + \eta_{ijt}$ and $\alpha_i \sim$ N(0,Σ) and $\eta_{it} \sim$ N(0,Ω). We can in principle simulate choice probabilities in this model just by drawing the α_i and η_{it} from the appropriate distributions and counting the frequency with which each option is chosen.

However, this approach is not practical, because if we draw the entire composite error $v_{ijt} = \alpha_{ij} + \varepsilon_{ijt}$ the model will deterministically generate particular choices and choice sequences, as in equations (6)-(7). So, equation (5) would become:

$$\hat{P}_{F,D}\left(d_{ij(1),1}, \dots, d_{ij(T),T} | X_{i1}, \dots, X_{iT}\right) = \frac{1}{D}\sum_{d=1}^{D}\prod_{t=1}^{T} I\left(d_{ij(t),t} | X_{it}, \alpha_d, \varepsilon_{dt}\right) \qquad (19)$$

where $I\left(d_{ij(t),t} | X_{it}, \alpha_d, \varepsilon_{dt}\right)$ is an indicator function for the choice $d_{ij(t),t}$ being observed at time t given the draws α_d and ε_{dt}. The practical problem is that the number of possible sequences is J^T. As we noted in the introduction, this is a very large number even for modest J and T. As a result, most individual sequences have very small probabilities. Hence, even for large D the value of (19) will often be zero. As Lerman and Manski (1981) discussed, very large simulations sizes are needed to provide accurate simulated probabilities of low probability events.

A potential solution to this problem, proposed by Berkovec and Stern (1991) and Stern (1992), is to reformulate the model to "add noise" and "smooth out" the indicator functions in (19). For instance, we could re-write the model as $U_{ijt} = \alpha_{ij} + X_{ijt}\beta + \varepsilon_{ijt} + \omega_{ijt}$, where all the serial correlation in the time-varying errors is captured by the ε_{ijt} process, while the ω_{ijt} are *iid* random variables (perhaps normal or extreme value). Then (7) is replaced by the condition:

$$U_{ijt} > U_{ikt}\ \forall\ k \neq j \ \Rightarrow\ \omega_{ijt} \geq -X_{ijt}\beta + \alpha_{dj} + \varepsilon_{djt} + (X_{ikt}\beta + \alpha_{dk} + \varepsilon_{dkt} + \omega_{ikt})\ \forall\ k \neq j \qquad (20)$$

These inequalities generate conditional probabilities $P\left(d_{ij(t),t} | X_{it}, \alpha_d, \varepsilon_{dt}\right)$ where P(·|·) depends on the distribution chosen for the ω_{ijt}. These probabilities are smooth functions of the model

parameters provided that the ω_{ijt} are continuous random variables. Simply plug these expressions into (19) to obtain:

$$\hat{P}_{SF,D}\left(d_{ij(1),1}, \ldots, d_{ij(T),T} | X_{i1}, \ldots, X_{iT}\right) = \frac{1}{D}\sum_{d=1}^{D}\prod_{t=1}^{T} P\left(d_{ij(t),t} | X_{it}, \alpha_d, \varepsilon_{dt}\right) \qquad (21)$$

Note that there are two ways to interpret (21). One could consider (20) the "true" model and view (21) as an unbiased probability simulator for this model. Alternatively, one could view the errors ω_{ijt} as simply a smoothing device, and view (21) as a smoothed version of (19). Such *ad hoc* smoothing will induce bias in the simulator, as noted by McFadden (1989).

The normal mixture of logits model (N-MIXL), where the α_{ij} are normal while the η_{ijt} are *iid* extreme value, can also be easily modified to accommodate serially correlated idiosyncratic shocks. Since the probability simulator for this model (equation (17)) is a frequency simulator, the procedure is exactly like what I just described, except in reverse. In this case the extreme value errors ω_{ijt}, which are present in the basic model, play the role of the "noise" that smooths the simulated probabilities. It is the serially correlated shocks ε_{ijt} that are added.

3.C. Extension to Include State Dependence

Finally, consider including true state dependence ($\gamma \neq 0$) in the model in (1)-(2). The difficult here is that we must not only simulate the error terms, but also lagged choices. Methods based on frequency simulation are not easily extended to this case. We can easily simulate entire choice histories from the model in (1)-(2) by drawing the α_i and η_{it} from the appropriate distributions. In each period these draws imply that one choice is optimal, as it satisfies (2). This choice is then treated as part of the history when we move on to simulate data for the next period. So the frequency simulator in (19) would become:

$$\hat{P}_{F,D}\left(d_{ij(1),1}, \ldots, d_{ij(T),T} | X_{i1}, \ldots, X_{iT}\right) = \frac{1}{D}\sum_{r=1}^{D}\prod_{t=1}^{T} I\left(d_{ij(t),t} | X_{it}, \alpha_r, \varepsilon_{rt}, d_{r,t-1}\right) \qquad (22)$$

where $I\left(d_{ij(t),t} | X_{it}, \alpha_r, \varepsilon_{rt}, d_{r,t-1}\right)$ is an indicator function for the choice $d_{ij(t),t}$ being observed at time t given the draws α_r and ε_{rt} and the lagged simulated choice $d_{r,t-1}$. The practical problem here is the same as we discussed in the $\gamma = 0$ case. The number of sequences is so large that we are unlikely to obtain draws that are consistent with a consumer's observed choice history, so in most cases (22) will simply be zero. Very large simulations sizes are needed to provide accurate simulated probabilities of low probability events.

In contrast, the GHK algorithm can be easily applied to estimate models that include individual effects, serial correlation and structural state dependence without any modification to the procedure described earlier. This is because the central idea of the algorithm is to construct random draw sequences that are required to be consistent with a consumer's observed choice history. These are then used to simulate transition probabilities from the choice at t-1 to the choice at t. (See equations (14)-(15) and the surrounding discussion).

In fact, Keane (1993, 1994) interpreted the GHK algorithm as an importance sampling algorithm where stochastic terms are drawn in a constrained way so that they must be consistent with observed choice histories (see equation (7)). These draws are not taken from the correct distribution given by $\alpha_i \sim N(0,\Sigma)$, $\eta_{it} \sim N(0,\Omega)$ and ρ. Rather, this is only used as a source density to generate draws that satisfy the constraints implied by the observed choice history. Importance sampling weights are then applied to these sequences when they are used to construct the probability simulator. That is, when taking the average over draws as in (14), sequences of draws that have greater likelihood under the correct distribution are given more weight. It turns out that in GHK the importance sampling weights simplify to transition probabilities as in (14).

There are ways to use frequency simulation in conjunction with smoothing or importance sampling to construct feasible simulators in the presence of state dependence. For example Keane and Wolpin (2001) and Keane and Sauer (2010) develop an algorithm based on the idea that all discrete outcomes are measured with some classification error. Then, any simulated draw sequence has a positive probability of generating any observed choice history. This is the probability of the set of misclassifications needed to reconcile the two histories. But this approach is not likely to be useful in most demand estimation contexts, as scanner data measuer choices quite accurately.

4. Testing for the Existence State Dependence

A large part of the literature on panel data discrete choice models of consumer demand has been concerned with estimating the degree of true state dependence in choice behavior. Researchers have been concerned with the question of whether, and to what extent, the observed (substantial) persistence in choice behavior over time can be attributed to unobserved individual effects and/or serially correlated tastes on the one hand, vs. true state dependence on the other.

We can gain some valuable intuition into the nature of state dependence by considering

the linear case. So we reformulate equation (1) to be:

$$U_{it} = \alpha_i + X_{it}\beta + \gamma U_{i,t-1} + \varepsilon_{it} \quad where \quad \varepsilon_{it} = \rho\varepsilon_{i,t-1} + \eta_{it} \tag{23}$$

where now U_{it} is an observed continuous outcome. I have suppressed the *j* subscripts to save on notation. By repeated substitution for the lagged U_{it}, we obtain:

$$U_{it} = \alpha_i\left(\frac{1-\gamma^t}{1-\gamma}\right) + \left(X_{it} + \gamma X_{i,t-1} + \cdots + \gamma^t X_{i1}\right)\beta + \gamma^{t+1}U_{ij0} + \left(\varepsilon_{it} + \gamma\varepsilon_{i,t-1} + \cdots + \gamma^t\varepsilon_{i1}\right) \tag{24}$$

Here U_{ij0} is the initial condition of the process. In conventional panel data analysis with large N and small T the treatment of initial conditions is often quite critical for the results. But in scanner data panels, where T is typically much larger, the results are not usually very sensitive to the treatment of initial conditions. Hence, I will not dwell on this topic here. Wooldridge (****) has an excellent discussion of this topic.

The critical thing to note about (24) is that lagged *X*s matter for the current *U* iff $\gamma \neq 0$. Thus, the key substantive implication of structural state dependence is that lagged *X*s help to predict current outcomes. This point was emphasized by Chamberlain (1984, 1985). But, as Chamberlain (1985, p.14) noted, "In order to make the distinction [between serial correlation and true state dependence] operational, there must be at least one variable which would not have a distributed lag response in the absence of state dependence." That is, to test for state dependence we need at least one variable X_{it}^k where we are sure that <u>lagged</u> X_{it}^k does not affect U_{it} directly, but only affects it indirectly through its affect on lagged *U*. This is analogous to saying that we have an X_{it-1}^k that is a valid instrument for $U_{i,t-1}$ in equation (23).

To be concrete, in consumer demand applications using scanner data, the covariates in *X* are typically (i) the observed characteristics of the products in the choice set, which are typically time invariant, (iii) a set of brand intercepts, which capture intrinsic preferences for brands and/or mean preferences for the unobserved attributes of brands, and (iii) the "marketing mix" variables, such as price, promotion activity and advertising activity, which are time varying. As only the marketing mix variables are time varying, at least one of these (price, display, ad exposures, etc.) must play the role of X_{it-1}^k in our effort to identify true state dependence.

Is it plausible that a variable like price would affect current demand only through its affect on lagged demand? At first glance the answer may seem completely obvious: Why should

the lagged price affect current demand? After all, it doesn't enter the consumer's current budget constraint. Isn't the only plausible story for why lagged price would predict current demand that it shifts lagged demand, which then affects current demand via some state dependence mechanism (like habit persistence, inventory, switching costs, etc.)?

But a closer examination of the issue reveals that there are subtleties. For example, Berry, Levinson and Pakes (1995) argue that prices of different car models may be positively correlated with their unobserved (*to the econometrician*) quality. This would tend to bias price elasticities of demand toward zero. They proposed using exogenous instruments for price to deal with this problem. Notably, however, they considered data with only one or a few periods. In the scanner data context, where there are many periods, it is much more straightforward to use brand intercepts to capture unobserved attributes of brands. In the typical scanner data context, once one controls for brand intercepts, there is no reason to expect that prices are correlated with unobserved attributes of the alternatives.

In contrast to the brand intercepts, which capture mean preferences for the unobserved attributes of products, the α_i are mean zero random variables which are interpreted as capturing heterogeneity in tastes for unobserved attributes of products. In my view, it is also plausible that prices are uncorrelated with the α_i. Why would the price of a product be correlated with person *i*'s intrinsic taste for that product? One person's tastes are too insignificant a part of total demand to affect the price of a product. In general, the random effects assumption:

$$E\left(\alpha_{ij}|X_{ij1},\ldots,X_{ijT}\right)=E\left(\alpha_{ij}\right)=0 \qquad (25)$$

is plausible when the *X*s include only brand attributes and marketing mix variables like price.

Finally, consider the time-varying taste shocks ε_{ijt}. It seems highly implausible that idiosyncratic taste shocks of individuals could affect the price of a product. Thus I would argue it is quite plausible that the strict exogeneity assumption holds:

$$E\left(\varepsilon_{ijt}|X_{ij1},\ldots,X_{ijT}\right)=0 \qquad (26)$$

But this assumes the ε_{ijt} are independent across consumers. A source of potential concern is aggregate taste shocks that generate cross-sectional dependence. But I would argue that, in weekly data, it is implausible that unanticipated aggregate taste shocks could influence the weekly price. In most instances there is simply not enough time for retailers to assess the demand

shift and alter prices so quickly. On the other hand, seasonal demand shocks are presumably anticipated long enough in advance to be reflected in the price are. Thus, I would argue that (26) is plausible even in the presence of aggregate shocks, provided one includes seasonal dummies.

These arguments support treating price and other marketing mix variables as strictly exogenous in (24), and estimating this equation by random effects. Let's say we assume that price is a "variable which would not have a distributed lag response in the absence of state dependence." Then we can test for the existence of state dependence by testing the significance of lagged price variables.

So far we have presented arguments that price is strictly exogenous with respect to idiosyncratic consumer tastes, but we have not yet confronted the question of whether lagged prices might have a direct effect on current demand U_{it}. In fact, there are a number of reasons to expect it might. I will describe three mechanisms that may generate such an effect:

(i) Reference price effects. There is a large literature in marketing arguing that consumer demand does not depend of price itself but rather on how the price compares to a "reference price." Key early work in this area was by Winer (1986). The reference price is typically operationalized as the average price of a product, or as some moving average of past prices. Reference price effects were originally motivated by psychological theories of choice. For instance, if the current price is higher than the reference price the consumer may perceive the price as "unfair" and be unwilling to pay it. But regardless of how one rationalizes the reference price variable, its existence implies that all lagged prices help to predict current demand.

(ii) Inventory effects. Erdem, Imai and Keane (2003) argued that reference price effects could be motivated as resulting from inventory behavior. If a product is storable, consumers will try to time their purchases for when price is relatively low. This creates an economic rationale for consumers to care about current price relative to a reference price. More generally, consumers are more likely to buy if current price is low relative to expected future prices. Thus, lagged prices matter if they are useful for forecasting future prices.

(iii) Price as Signal of Quality. Another mechanism for lagged prices to have a direct effect on current demand is if consumers have uncertainty about product attributes and use price as a signal of quality. Erdem, Keane and Sun (2008) estimated a model of this form. In such a model, a history of high prices will cause relatively uninformed consumers to infer that a brand is high quality. As a result, willingness to pay for a product is increasing in its own lagged prices.

Of course, such a mechanism becomes less important as consumers gain experience with a product category.

In all the above examples, the true model exhibits some form of dynamics, but not what is generally known as true state dependence. As Chamberlain (1985, p.12) states, "The intuitive notion is that if occupancy of a state affects an individual's preferences or opportunities, then there is state dependence." This intuitive notion does not hold in the above three examples: (i) in the reference price model the actual purchase of a brand has no effect on its reference price. Only price realizations affect references prices. (ii) In the inventory model, lagged prices of a brand only matter because they affect expected future prices.[3] (iii) The signaling model resembles the reference price model in that higher lagged prices increase willingness to pay for a brand.

Conversely, there are plausible cases where lagged prices are insignificant in (24) but true state dependence nevertheless exists. A well-known class of structural models that generates true state dependence is the consumer learning model. In the learning model consumers have uncertainty about product attributes and learn about them over time through use experience, advertising and other signals. Examples of structural learning models are Ecsktein, Horsky and Raban (1988), Roberts and Urban (1988), Erdem and Keane (1996), Ackerberg (2003), Crawford and Shum (2005) and Ching (2010). In the learning model of Erdem and Keane (1996), which Keller (2002) calls "the canonical *economic* model of brand equity," consumers are risk averse with respect to variability in brand quality. As a result, they are willing to pay a premium for familiar brands whose quality is relatively certain, as opposed to less familiar brands with equal expected quality but greater uncertainty. For this reason, lagged purchases affect the current utility evaluation, because they reduce ones uncertainty about a product's attributes.

Thus, if we estimate (24), and the true model is a learning model, we might expect to find that lagged prices matter because they influence lagged purchase decisions. But this is not so clear. In the simplest Bayesian learning model, with use experience as the only signal, the perceived variance of brand *j* at time *t* is:

$$\sigma_{ijt}^2 = \left[\left(1/\sigma_{ij0}^2\right) + N_{ij}(t)(1/\sigma_{\varepsilon}^2)\right]^{-1} \qquad (27)$$

[3] Of course, lagged purchases do affect current inventory, which is a state variable. And, if there are inventory carrying costs, consumers are less likely to buy, *ceteris paribus*, if current inventory is high. Furthermore, current inventory is more likely to be high in cases where recent lagged prices were low. But note that inventory is affected by past purchase of a *category*, not purchase of a particular *brand*.

Here, σ_{ij0}^2 is consumer i's prior uncertainty about the quality of brand j, while σ_ε^2 is variability of experience signals. $N_{ij}(t)$ is the total number of times that consumer i bought brand j prior to t. We would expect lower lagged prices to lead to higher $N_{ij}(t)$ and hence lower σ_{ijt}^2. But, at the same time, a brand with relatively low σ_{ijt}^2's (across all consumers in the market) may charge relatively high prices because it has more brand equity. This leaves the correlation between lagged prices and current demand ambiguous.

This argument amounts to a statement that estimates of (24) may be unrevealing because prices and the σ_{ijt}^2 are jointly determined in the learning model – rendering prices endogeneous in (24). Fully structural estimation of the learning model resolves this problem by modeling the relationship between prices, the $N_{ij}(t)$ and the σ_{ijt}^2. But of course this requires a strong set of maintained structural assumptions.

In light of the above arguments, I do not believe that the significance or insignificance of prices (or other marketing mix variables) in (24) provides a relatively "assumption free" test of whether true state dependence exists. If lagged prices are significant, it may be because of reference price, inventory, quality signaling or other factors that cause lagged prices to directly influence current demand. Conversely, insignificance of lagged prices does not necessarily rule out the existence of state dependence, as illustrated by the example of the learning model.

Now consider the additional issues that arise in testing for state dependence in the case of a discrete dependent variable, as in (1)-(2). Recall from our discussion in Section 3, that in the case of a random effect but no state dependence (or other forms of serial correlation), we have:

$$P\left(d_{it}|d_{i1}, d_{i2}, \dots, d_{i,t-1}, X_{i1}, X_{i2}, \dots, X_{i,t-1}, X_{it}\right) \neq P(d_{it}|X_{it}) \qquad (28)$$

Thus, the choice probability at time t depends on the whole history of the process $\{d_{is}, X_{is}\}_{s=1}^{t-1}$, and not just on X_{it}. In equation (10), we gave a simple intuition for why, based on a three period case with only two alternatives, where the consumer chooses option 2 in all three periods:

$$P(d_{i3} = 2|d_{i1} = 2, d_{i2} = 2, X_{i1}, X_{i2}, X_{i3}) = P(d_{i3} = 2|\nu_{i1} \geq -X_{i1}\beta, \nu_{i2} \geq -X_{i2}\beta, X_{i3}) \qquad (10')$$

That is, the reason the whole past history helps to predict d_{it} is that we can't observe lagged utility, only lagged choices. But information on lagged choices, such as d_{i1}=d_{i2}=2, implies conditions like $\nu_{i1} \geq -X_{i1}\beta$ and $\nu_{i2} \geq -X_{i2}\beta$, which are informative about the distribution of

the current error. In fact, as we noted earlier, the conditional density of v_{i3} in this case has the form $f(v_{i3} \mid v_{i1} \geq -X_{i1}\beta, v_{i2} \geq -X_{i2}\beta)$. This exact same argument holds regardless of whether the source of serial correlation in the errors is a random effect, serial correlation in the time-varying error component, or both.

As Heckman (1981) discussed, the fact that lagged choices help to predict the current error means that lagged choices will tend to be significant in a discrete choice model with serial correlation, even if there is no true state dependence. This phenomenon is known as "spurious state dependence." The fact that the whole history matters when there is serial correlation makes it extremely difficult to distinguish true state dependence from serial correlation.[4]

Nevertheless, an important positive result about identification in the probit model is the following: Assume that the errors η_{ijt} in (1) are normal, and that $\alpha_i \sim N(0,\Sigma)$, giving a random effects probit. Then the coefficient γ on the lagged dependent variable is identified in (1). This is because, as Chamberlain (1984, p.1279) notes: "the most general multivariate probit model cannot generate a Markov chain. So we can add a lagged variable and identify γ." That is, if the multivariate distribution of the composite errors $v_i = \{v_1,\ldots,v_T\}$ is diagonal (no serial correlation), the probit (with γ=0) generates that choices are independent over time (conditional on X_i). Alternatively, if the errors are serially correlated (but γ=0) then the whole history of choices prior to time t helps to predict the choice at time t. The intermediate case of a Markov process cannot be attained regardless of the specification of the error structure. It can only be attained by including a lagged dependent variable (i.e., allowing $\gamma \neq 0$).

There are two practical implications of these results:

First, if one estimates a discrete choice model without adequately controlling for random effects and serial correlation, then one is likely to find spurious state dependence. Indeed, numerous studies since Guadagni and Little (1983) have found that the estimated strength of state dependence in consumer brand choices declines substantially when one controls for heterogeneity and serial correlation.

Second, within the probit framework, one can test if state dependence exists by including rich controls for heterogeneity and serial correlation and then testing the significance of lagged

[4] Note that a random effect will generate a situation where all lags are equally informative about the current error term. In contrast, a process like a stationary AR(1) generates a situation where more recent choices are more informative, although the whole history still matters. Even if the errors are MA(1), the whole history of the process helps to predict the current choice.

dependent variables. This approach was pursued in Keane (1997) and in a number of subsequent papers, such as Paap and Franses (2000), Smith (2005), Dubé, Hitsch and Rossi (2010) and many others. This work consistently finds evidence for the existence of state dependence.

Chamberlain argued, however, that tests within the probit framework were suspect because of their reliance on the probit functional form – in particular, the fact that it is not possible within the probit framework to choose and error structure that generates a Markov chain. Chamberlain (1985, p.14) went on to suggest that a test based on regressing the current choice on current and lagged *X*s (and controlling for heterogeneity) "should not be very sensitive to functional form." However, we discussed tests based on lagged *X*s (especially price) earlier, and found that strong economic assumptions underlie such tests in the consumer demand context.

Chamberlain (1985) went on to argue that a completely non-parametric test for state dependence cannot exist, because one can always find a latent variable α_i such that:

$$P(d_{it}|X_{i1},\dots,X_{iT},\alpha_i) = P(d_{it}|X_{it},\alpha_i) = P(d_{it}|\alpha_i) \tag{29}$$

That is, one can always find a distribution of α_i such that $\{d_{i1},\dots,d_{iT}\}$ is independent of $\{X_{i1},\dots,X_{iT}\}$. He gives a simple example (p. 1281) where α_i is simply a unique integer assigned to every different configuration of *X*s in the data. This is equivalent to a latent class model with a discrete distribution of types. Each type has its own vector of multinomial choice probabilities. And each configuration of *X*s in the data corresponds to a different type. Then, type summarizes all the information in the *X*s, giving independence of *d* and *X* conditional on α.

Chamberlain defines a relationship of *X* to *d* as "static" conditional on α if *X* is strictly exogenous (conditional on α) and if d_t is independent of $\{X_{i1},\dots,X_{i,t-1}\}$ conditional on X_t and α. If a relationship is static there is no structural state dependence. Equation (29) implies there <u>always</u> exists a specification of α such that the relationship of *X* to *d* is static. Thus, we cannot test for structural state dependence without imposing some structure on $P(\cdot|\cdot)$ and the distribution of α.

However, I do not view this negative result as disturbing. As Koopmans et al (1950) noted long ago, we cannot learn anything of substance from data without making some *a priori* structural assumptions (see Keane (2010a,b) for discussion of this issue). So I would be very surprised if that were not true with regard to drawing inferences about state dependence. In other words, the fact that our inferences about the nature of state dependence, heterogeneity and serial

correlation in tastes are contingent on our modeling assumptions is not at all unique to this set of issues. It is the normal state of affairs throughout economics and the natural sciences as well.[5]

A good example of imposing structure is Chamberlain (1984)'s "correlated random effects probit model," henceforth CRE. In this model, α_i is constrained to be a linear function of the time varying elements of X_i, which I denote by Z_i, plus a normal error term, giving:

$$\alpha_{ij} = Z_{ij1}\delta_{j1} + \cdots + Z_{ijT}\delta_{jT} + \mu_{ij} \tag{30}$$

Note that the effect of time-invariant elements of X_i on α_i is not identified separately from the intercepts; letting a time-invariant element of X_i shift α_i would be equivalent to letting it shift $X_{it}\beta$ by a constant. Given (29), one can test for state dependence and strict exogeneity.

A CRE model combining (1)-(2) with (30) may be very useful if the *X*s are individual characteristics, which obviously may be correlated with preferences (see Hyslop (1999) and Keane and Sauer (2010) for recent labor applications). But in the consumer demand context, the *X*s are not usually characteristics of people but rather of products, including marketing variables like price and advertising. Here, I think the CRE model is not very compelling.

In particular, I argued earlier that a standard random effects assumption on α_i is plausible in the consumer demand context (see equation (25)). The most obvious time-varying attribute of a product is price. It is clearly implausible that price would be affected by individual brand preferences. But before ruling out correlation between α_i and price we should also ask, "What is the source of price variation in prices across consumers and over time?" Erdem, Imai and Keane (2003) argue that almost all price variation in scanner data is exogenous from the point of view of consumers. Pesendorfer (2002) and Hong et al (2002) argue that a type of inter-temporal price discrimination strategy on the part of firms, where retailers play mixed strategies, most plausibly

[5] Chamberlain (1985)'s negative results on non-parametric identification of state dependence do raise some interesting methodological questions. I will not attempt to address them here, but it is worth raising them: (i) Chamberlain allows for extraordinarily general patterns of heterogeneity. Does Occam's razor (or just common sense modeling practice) suggest limiting ourselves to much more parsimonious forms like (25) or (30)?, (ii) It is not clear how a model where α_i is allowed to depend in a very general way on time varying *X*'s can be used for forecasting. Should we limit ourselves to more parsimonious models in the interest of forecasting ability?, (iii) In light of Chamberlain's negative results, and our own discussion surrounding equation (24), should we conclude that state dependence is not a useful construct in demand modeling? Would it be more fruitful to focus directly on modeling the dynamics of how lagged *X*s affect current and future choices, without the mediating concept of state dependence?, (iv) Alternatively, is the state dependence construct useful because it enables us to develop more convenient and parsimonious functional forms compared to including many lagged covariates in a model?

explains the frequent week-to-week price fluctuations for frequently purchased consumer goods that we see in scanner data.[6] This price variation would appear random to consumers.

In light of these observations, I would place considerable confidence in results in the marketing and IO literatures that find substantial evidence of state dependence in consumer choice behaviour (provided the studies in question include adequate controls for consumer heterogeneity and serial correlation in tastes). The existence of state dependence is important, as it implies that current marketing actions, such as price discounts, affect not only current but also future demand. But an even more important question is what mechanism generates state dependence. I turn to this question in the next section.

5. Empirical Work and State Dependence and Sources of Dynamics in Demand

In this section I discuss attempts to identify and quantify sources of state dependence, and choice dynamics more generally. The field of marketing has reached rather broad consensus on many key issues related to the dynamics of consumer demand over the past 20 years, as I discuss below. The potential explanations for state dependence include learning, inventories and/or reference prices, habit persistence, variety seeking and switching costs. All of these have been examined, but learning and inventories have received the most attention in the literature. I'll start by discussing some of the more influential work on the functional form of state dependence.

After Guadagni and Little (1983), the main approach to modeling state dependence in the marketing literature was to let current utility depend on an exponentially smoothed weighted average of lagged purchase indicators, denoted GL_{ijt}. Specifically, replace $d_{ij,t-1}$ in (1) with:

$$GL_{ijt} = \theta GL_{ij,t-1} + (1-\theta)d_{ij,t-1} = (1-\theta)\left\{\sum_{s=1}^{t-1} \theta^{s-1}\, d_{ij,t-s} + \theta^{s-1} GL_{ij1}\right\} \qquad (31)$$

Guadagni and Little famously called GL_{ijt} the "brand loyalty" variable. The smoothing parameter

[6] There are sensible arguments for why consumer types may be correlated with brand prices, but I do not believe they are empirically relevant. Scanner data is typically collected from all the (large) stores in a particular area, like Sioux Falls, SD or Springfield, MO. So regional variation is not a potential source of price variation, but cross-store variation potentially is. However, while it is likely that stores differ in their average price level (e.g., some stores are more "up-scale," or are located in wealthier areas, and therefore charge higher prices in general), it not clear why *relative* prices of brands would differ by store. Another idea is that consumers may actively seek out stores where their preferred brand is on sale. Or, even if they regularly visit only one store, to time visits for when that store is having a sale on their preferred brand. Such behavior *might* be relevant for expensive goods (e.g., meat, wine, diapers), but I doubt that anyone would decide when or what store to visit based on the price of Oreo cookies. Some years ago I attempted (in joint work with Tulin Erdem) to develop a model of store choice based on prices of various items. But we abandoned the project as we could not find any products that predicted store choice.

$\theta \in [0,1)$ determines how quickly the impact of lagged purchases on current utility decays. If θ=0 then only $d_{ij,t-1}$ matters and we are back to a first order Markov process as in (1). As $\theta \rightarrow 1$ we get substantial inertia in brand preferences. For typical panel lengths and reasonable values of θ the initial setting of GL_{ij1} is not very important.

Guadagni and Little (GL) estimated their model using scanner data on coffee purchases of 100 households in Kansas City for 32 weeks in 1979. They estimated a MNL model with 8 alternatives. But they had no controls for heterogeneity or serial correlation in preferences (as this was not technically possible in 1983). Their complete model implied that "brand loyalty," along with price and promotional activity, are strong predictors of brand choice.

Keane (1997) considered the impact of allowing for random effects and AR(1) errors in a model with the GL form of state dependence. The data cover 51 weeks of ketchup purchases by 1,150 consumers in Sioux Falls, SD in 1987-88. The choice set contained 7 alternatives,[7] and up to 30 purchases per household. Thus, the required order of integration for the model with AR(1) errors is T(J-1) = 180, and choice probabilities were evaluated using the GHK algorithm.

Keane assumed that $\alpha_i \sim N(0,\Sigma)$ and $\eta_{it} \sim N(0,\Omega)$, giving a multinomial multi-period probit model. A major problem is that unrestricted Σ and Ω would contain T(J-1)J/2 – 1= 631 parameters. To deal with this, he assumed that both Σ and Ω had a one factor structure. Then the covariance structure is characterized by (i) the AR(1) parameter ρ, (ii) the 6 factor loadings on the common factor that underlies Σ, (iii) the uniquenesses of Σ, which are assumed equal for all brands and denoted by κ, and (iv) the same 7 parameters for Ω. This gives only 15 parameters. Although this structure is very parsimonious, additional factors were not significant.

One goal of Keane (1997) was to give a taxonomy of types of heterogeneity. He argued that to rationalize the most general models in the literature on needs 7 types: (i) observed and unobserved heterogeneity in tastes for observed attributes, (ii) observed heterogeneity in brand intercepts, (iii) unobserved heterogeneity in tastes for unobserved common and unique attributes for which consumers have fixed tastes, and (iv) the same for attributes where consumers have time varying tastes. The basic strategy in Keane (1997) was to add more and more types of heterogeneity and see how estimates of state dependence were affected.

[7] These were Hunt's (32 oz), Del Monte (32 oz), and five sizes of Heinz (40, 64, 14, 28, and 32 oz). For Heinz the 32 and 14 oz were glass and the other sizes were plastic. The Heinz 40-ounce size was introduced during in the sample, creating a nice source of variation in the choice set. Heinz 32 oz is the base alternative whose utility is normalized to zero.

Keane's "Model 1" is very similar to Guadagni and Little (1983) but with normal errors (panel probit). He estimates θ=.813 and $\lambda = 1.985$. Note that $\lambda(1-\theta) = .37$ is the extra utility from buying brand *j* at *t* if you bought it at *t*-1. The estimate of the price coefficient is -1.45, so this is equivalent to 27 cent price cut. As mean price is roughly $1.20, this is about a 22.5% price cut.

Keane's "Model 2" eliminates state dependence but includes heterogeneity in brand intercepts of the form $\alpha_i \sim N(0,\kappa I_{J-1})$. So we have unique factors but no common factors. The unique factors account for 48% of total error variance, imply substantial heterogeneity in tastes.

Keane's "Model 3" includes both the GL form of state dependence and "κ-heterogeneity" (i.e., unique factors). When both are include, each becomes less important. The fraction of the error variance due to unique factors drops to 31%. We now get θ=.833, $\lambda = 0.889$, and a price coefficient of -1.66. So the effect of lagged purchase is equivalent to only a 9 cent price cut.

In the full model ("Model 16"), which includes all 7 types of heterogeneity, $\lambda = 1.346$ and $\theta = .909$. The price coefficient is heterogeneous, but for a typical family it is about -2.4. So lagged purchase has an effect on demand that is similar to roughly a 5-cent price cut (4%).

The effect of a purchase today on the probability of a purchase tomorrow is known as the "purchase carry-over effect" in marketing. The bottom line of Keane (1997) is that extensive controls for heterogeneity reduce the estimated carry-over effect from being equivalent to a 22.5% price cut to a 4% price cut – thus reducing it by roughly 80%. So most of the observed persistence in brand choice does appear to be due to taste heterogeneity, but there is still a significant fraction that is due to state dependence.[8]

Of course, as we discussed in Section 4, inferences about the relative importance of heterogeneity and state dependence are always functional form dependent. Erdem and Keane (1996) showed that a Bayesian learning model implies a very different form of state dependence from that in GL. In their model, prior to receiving any information, consumers perceive that the true quality of brand *j*, denoted Q_j, is distributed normally with mean Q_{j0} and variance σ_{j0}^2. Over time a consumer receives noisy information about a brand through use experience and ad signals. Let d_{jt} be an indicator for whether brand *j* is bought at time *t*, and let σ_ε^2 denote the noise in

[8] Short run vs. long run price elasticities of demand are also of interest. In model 1 a 50% price cut leads to 257% sales increase in current period (elasticity of 5.1) but only about a 17% sales increase in subsequent periods (elasticity of roughly 0.34). In model 16 a 50% price cut leads to 313% sales increase in current period (elasticity of 6.3) but only about a 12% sales increase in subsequent periods (elasticity of roughly 0.24).

experience signals. Let d_{jt}^{A} be an indicator for whether an ad for brand *j* is seen at time *t*, and let σ_A^2 denote the noise in ad signals. Let $N_j(t)$ and $N_j^A(t)$ denote the total number of experience and ad signals received up through time *t*, respectively. Then the Bayesian learning model implies:

$$Q_{jt} = \frac{(1/\sigma_\varepsilon^2)}{\left(1/\sigma_{jt}^2\right)} \sum_{s=1}^{t-1} Q_{js}^{E}\, d_{js} + \frac{(1/\sigma_A^2)}{\left(1/\sigma_{jt}^2\right)} \sum_{s=1}^{t} A_{js}\, d_{js}^{A} + \frac{(1/\sigma_{j0}^2)}{\left(1/\sigma_{jt}^2\right)} Q_{j0} \tag{32}$$

$$\sigma_{jt}^2 = \frac{1}{\left(1/\sigma_{j0}^2\right) + N_j(t)(1/\sigma_\varepsilon^2) + N_j^A(t)(1/\sigma_A^2)} \tag{33}$$

Here, Q_{jt} is the perceived quality of brand *j* based on information received up through time *t*, and σ_{jt}^2 is the perception error variance.

Note that the Bayesian learning model implies a very different form of state dependence than GL. First, note that more lagged purchases ($N_j(t)$) reduce perceived uncertainty about the quality of a brand (σ_{jt}^2). If consumers are risk averse with respect to quality variation, this makes familiar brands more attractive, generating state dependence. The Bayesian framework in (33) implies that only the <u>total</u> number of lagged purchases of a brand, $N_j(t)$, matters for its current demand, while the GL framework in (31) implies that more recent experience is more important.

A more subtle difference between the models is that, in the learning model, heterogeneity and state dependence are not neatly separable phenomena. In (32), perceived quality of brand *j* at time *t*, Q_{jt}, is a function of all quality signals received up through *t*. This is heterogeneous across consumers – some will, by chance, receive better quality signals than others. Thus, heterogeneity in brand preferences evolves through time via the same process that generates state dependence.

Because the Q_{jt} are serially correlated random variables, which depend on lagged signals, we must use simulation to approximate the likelihood. What we have is a very complex mixture of logits model, with the mixing distribution given by the distribution of the Q_{jt}. The method used to simulate the likelihood is a smooth frequency simulator, like that presented in equation (21), with the ε_{dt} playing the role of the draws for the Q_{jt}.

Erdem and Keane (1996) compared a Guadagni and Little (1983) style model with a Bayesian learning model where state dependence is governed by (32)-(33).[9] They used Nielsen

[9] Erdem and Keane estimated two versions of their model where consumers are either myopic or forward-looking. Here I discuss only the myopic version, which is very similar to GL except for the different form

scanner data on liquid detergent purchases of 167 households in Sioux Falls, SD for 51 weeks in 1987-88. Telemeters were attached to panelists' TVs to measure ad exposures. The data include 7 brands, and a no purchase option. Three brands were introduced during the period, generating variability in brand familiarity. EK augment the Guadagni-Little model by including a GL-type variable for ad exposures. Thus, both past use experience and ad exposures affect current utility.

When Erdem and Keane estimated the GL model they obtained θ = .770 and λ = 3.363, so $\lambda(1-\theta)$ = .773. The price coefficient was -1.077, implying that the impact of lagged purchase is equivalent to roughly a 72 cent price cut. Mean price is roughly $3.50, so this is 21%. This is very close to the effect Keane (1997) found for the GL model for ketchup. Surprisingly, the λ for advertising was only 0.14 with a standard error of .31 (not significant). Thus, the GL model implies the awkward result that advertising has no effect on demand.

However, Erdem and Keane (1996) found the Bayesian learning model gave a much better fit to the data than the Guadagni-Little model. The log likelihood (LL) and Bayes Information Criterion (BIC) for the GL model were -7463 and 7531. But for the learning model they obtained LL and BIC values of -7312 and 7384. Thus, the BIC improvement is 147 points. The key parameters that generate state dependence are σ_{j0}^2=0.053, σ_ε=0.374 and σ_A=3.418.

The EK model is too complex to give simple calculations of the impact of lagged choices on current demand as we did with the GL and Keane (1997) models. The effects of price changes and changes in ad exposure frequency can only be evaluation by simulating the model. Unfortunately, EK only report advertising and not price simulations. But they do find clear evidence of state dependence in the advertising simulations. As they state, "although the short run effect of advertising is not large, advertising has a strong cumulative effect on choice over time as it gradually reduces the perceived riskiness of a brand."[10]

Based on the evidence in Erdem and Keane (1996) and Keane (1997), as well as a large body of subsequent work, much of which is very well described by Neslin (2002), there is now a broad consensus on three issues: (i) state dependence in demand does exist, (ii) as a result, both price promotion and advertising have long run effects, but (iii) consumer taste heterogeneity is a

---

of state dependence. The myopic model can be estimated using methods discussed in Section 3. The forward-looking version requires dynamic programming, which is beyond the scope of this paper.

[10] Unfortunately, their paper contains a major typo in key figure (Figure 1) that shows this result. The Figure 1 in the paper just duplicates Figure 3. Fortunately, the basic result can also be seen in Figure 2 (for the model with forward-looking consumers).

much stronger source of the observed persistence in choice behavior than is state dependence.

In contrast to the consensus on existence of state dependence, there is no clear consensus on its source. The Guadagni and Little (1983) and Keane (1997) types of model can be viewed as structural models where prior use experience literally increases the utility of current consumption of a brand through a habit persistence mechanism. Alternatively, these models can be viewed as flexible approximations to a broad (but unspecified) set of models that generate state dependence that is well described by the "brand loyalty" variable. The Erdem and Keane (1996) model and the large body of subsequent work derived from it (see Ching, Erdem and Keane (2013) for a review) definitively takes stand that state dependence derives from the learning mechanism. Other work, especially Erdem, Imai and Keane (2003) and Hendel and Nevo (2006), posits that inventories are an importance source of dynamics. Erdem, Keane and Sun (2008) show that the learning and inventory mechanisms are actually very hard to disentangle empirically, if one allows for *a priori* consumer taste heterogeneity. There is little consensus on the relative importance of the different mechanisms that may generate state dependence.

The third key research objective that I mentioned in the introduction is to understand the dynamics of demand. Most important is to understand the sources of the observed increase in demand when a brand is on sale. Here, I think the literature has reached a high degree of consensus. Consider the demand for frequently purchased consumer goods. There is broad consensus that own price elasticities (given temporary price cuts) are about –3 to –4.5. But it is also widely accepted by firms and academics just knowing how much demand goes up when you cut prices is not very interesting. What really matters is where the increase comes from.

Erdem, Imai and Keane (2003) and Erdem, Keane and Sun (2008) estimate that roughly 20-30 percent of the increase in sales due to a temporary price cut is cannibalization of future sales. Of the remaining incremental sales, 70-80 percent is due to category expansion and only about 20-30 percent is due to brand switching. It is hard to exaggerate the importance of this 3-way decomposition of the price elasticity of demand, as it determines the profitability of price promotion. And a remarkable consensus has emerged on these figures in recent years. Some key papers on cannibalization rates are van Heerde, Leeflang and Wittink (2000, 2004) and Ailawadi, Gedenk, Lutzky, Neslin (2006). And some important studies of brand switching are Pauwels, Hanssens and Siddarth (2002), van Heerde, Gupta and Wittink (2003), Sun, Neslin and Srinivasan (2003) and Mace and Neslin (2004).

6. Conclusion

As we have seen, there is broad consensus that state dependence in consumer demand exists. There is also clear evidence that dynamic demand models fit the data much better than static models (see Ching, Erdem and Keane (2008)). And there is broad agreement that only about 20-25% of the incremental sales that accompany a price cut is due to brand switching, with the rest due to category expansion and cannibalization of own future sales. On the other hand, there is little agreement on the fundamental mechanism that generates dynamics in demand. The main competing theories are learning, inventories and habit persistence. Progress in this area is severely hindered by the computational difficulty of nesting all this mechanisms in one model.

Much of demand modeling is done with the ultimate goal of merging the demand side with supply side models of industry competition. Such equilibrium models can be used for merger analysis, advertising regulation, anti-competitive pricing regulation, etc. But existing work in this area has typically used static demand models, due to the computational difficulty of solving the problem of oligopolistic firms when demand is dynamic.

Unfortunately, static demand models greatly exaggerate cross-price elasticities, as they attribute too much of incremental sales to switching (see Sun, Neslin and Srinivasan (2003) and Erdem, Imai and Keane (2003), Erdem, Keane and Sun (2008)). As cross-price elasticities of demand summarize the degree of competition between products, this bias will create serious problems in attempting to predict effects of mergers. This example that makes obvious the importance of further work on developing dynamic models, particularly ones that are sophisticated enough to capture observed dynamics, yet simple enough to merge with supply side models.

References

Ackerberg, D. (2003) Advertising, learning, and consumer choice in experience good markets: A structural empirical examination. *International Economic Review*, 44(3): 1007-1040.

Ailawadi, K., K. Gedenk, C. Lutzky and S. Neslin (2007),"Decomposition of the sales impact of promotion-induced stockpiling," *Journal of Marketing Research*, 44:3, 450-467

Allenby, G. M., & Lenk, P. J. (1994). Modeling household purchase behavior with logistic normal regression. *Journal of American Statistical Association, 89*, 1218-1231.

Andrews, R.L. and A.K. Manrai (1999), "MDS Maps for Product Attributes and Market Response: An Application to Scanner Panel Data," *Marketing Science*, 18(4), 584-604.

Berkovec, James & Stern, Steven, 1991. "Job Exit Behavior of Older Men," *Econometrica*, Econometric Society, vol. 59(1), pages 189-210, January.

Berry, Steven (1994), "Estimating Discrete Choice Models of Product Differentiation," RAND Journal of Economics, 25, 242-262.

Berry, S., J. Levinsohn, and A. Pakes. (1995). ''Automobile Prices in Market Equilibrium'', Econometrica 63, 841–890.

Block, H. andi. Marschak (1960), "Random Orderings and Stochastic Theories of Response," in I. Olkin, ed., *Contributions to Probability and Statistics*, Stanford University Press.

Burda, M., M. Harding and J. Hausman (2008), A Bayesian mixed logit-probit model for multinomial choice. *Journal of Econometrics* 147: 232-246.

Butler, J S & Moffitt, Robert, 1982. "A Computationally Efficient Quadrature Procedure for the One-Factor Multinomial Probit Model," *Econometrica*, 50(3), pages 761-64, May.

Chamberlain, Gary (1984). "Panel Data," in Handbook of Econometrics, Volume 2, eds. Z. Griliches and M. Intriligator, Amsterdam: North-Holland, pp. 1247–1318.

Chamberlain, Gary (1985). "Heterogeneity, Omitted Variable Bias, and Duration Dependence," in Longitudinal Analysis of Labor Market Data, eds. J. Heckman and B. Singer, Cambridge: Cambridge University Press, pp. 3–38.

Ching, A.T. (2010) Consumer learning and heterogeneity: dynamics of demand for prescription drugs after patent expiration. *International Journal of Industrial Organization*, 28(6): 619-638.

Ching, A., T. Erdem and M. Keane (2009). The Price Consideration Model of Brand Choice, *Journal of Applied Econometrics*, 24:3, 393-420.

Ching, A., T. Erdem and M. Keane (2013). Learning Models: An Assessment of Progress, Challenges and New Developments." *Marketing Science*, forthcoming.

Crawford, G. and M. Shum (2005). Uncertainty and learning in pharmaceutical demand. *Econometrica*, 73(4): 1137–1173.

Jean-Pierre Dubé & Günter J. Hitsch & Peter E. Rossi, 2010. "State dependence and alternative explanations for consumer inertia," *RAND Journal of Economics*, vol. 41(3), pages 417-445.

Elrod, Terry. (1988). ‘‘Choice Map: Inferring a Product Map from Observed Choice Behavior’’, Marketing Science 7 (Winter), 21–40.

Elrod, T. and M. Keane. (1995). ‘‘A Factor-Analytic Probit Model for Representing the Market Structure in Panel Data’’, *Journal of Marketing Research*, 32, 1–16.

Erdem, T. and M. Keane (1996). "Decision Making under Uncertainty: Capturing Dynamic Brand Choice Processes in Turbulent Consumer Goods Markets," *Marketing Science*, 15:1, 1-20.

Erdem, T., S. Imai and M. Keane (2003). "Brand and Quantity Choice Dynamics under Price Uncertainty," *Quantitative Marketing and Economics*, 1:1, 5-64.

Erdem, T., M. Keane and B. Sun (2008). "A Dynamic Model of Brand Choice when Price and Advertising Signal Product Quality," *Marketing Science*, 27:6, 1111-25.

Geweke, J. and M. Keane (1999), Mixture of Normals Probit Models. in *Analysis of Panels and Limited Dependent Variable Models*, Hsiao, Lahiri, Lee and Pesaran (eds.), Cambridge University Press, 49-78.

Geweke, J. and M. Keane (2001), Computationally Intensive Methods for Integration in Econometrics. In *Handbook of Econometrics: Vol. 5*, J.J. Heckman and E.E. Leamer (eds.), Elsevier Science B.V., 3463-3568.

Geweke, J., Keane, M. and D. Runkle (1994). Alternative Computational Approaches to Statistical Inference in the Multinomial Probit Model. Review of Economics and Statistics, 76:4, 609-32.

Geweke, J., Keane, M. and D. Runkle (1997). Statistical Inference in the Multinomial Multiperiod Probit Model. Journal of Econometrics, 80, 125-65.

Guadagni, Peter M. and John D.C. Little. (1983). ‘‘ALogit Model of Brand Choice Calibrated on Scanner Data’’, Marketing Science 2 (Summer), 203–238.

Hajivassiliou, Vassilis & McFadden, Daniel & Ruud, Paul, 1996. "Simulation of multivariate normal rectangle probabilities and their derivatives theoretical and computational results," *Journal of Econometrics*, Elsevier, vol. 72(1-2), pages 85-134.

Harris, K. and M. Keane (1999), "A Model of Health Plan Choice: Inferring Preferences and Perceptions from a Combination of Revealed Preference and Attitudinal Data," *Journal of Econometrics*, 89: 131-157.

Heckman, J.J. (1981) Heterogeneity and State Dependence. In S. Rosen (ed.), *Studies in Labor Markets*: 91-140.

Hendel, I. and A. Nevo (2006) Measuring the Implications of Sales and Consumer Inventory Behavior. *Econometrica*, 74(6): 1637-73.

Hong, Pilky, R. Preston McAfee and Ashish Nayyar. (2002). ''Equilibrium Price Dispersion with Consumer Inventories,'' *Journal of Economic Theory* 105, 503–517.

Dean R. Hyslop, 1999. "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women," *Econometrica*, 67(6), pages 1255-1294.

Kamakura, Wagner and Gary Russell (1989), "A Probabilistic Choice Model for Market Segmentation and Elasticity Structure," *Journal of Marketing Research*, 26, 379-390.

Keane, M. (1992). A Note on Identification in the Multinomial Probit Model. Journal of Business and Economic Statistics, 10:2, 193-200.

Keane, Michael P. (1993). ''Simulation Estimation for Panel Data Models with Limited Dependent Variables''. In G.S. Maddala, C.R. Rao, and H.D. Vinod (eds.), Handbook of Statistics II: Econometrics. Amsterdam: Elsevier Science Publishers.

Keane, Michael P. (1994). ''A Computationally Practical Simulation Estimator for Panel Data'', Econometrica 62(1), 95–116.

Keane, Michael P. (1997). ''Modeling Heterogeneity and State Dependence in Consumer Choice Behavior'', Journal of Business and Economic Statistics 15(3), 310–327.

Keane, M. (2010a) Structural vs. Atheoretic approaches to econometrics. *Journal of Econometrics*, 156(1): 3-20.

Keane, M. (2010b) A Structural Perspective on the Experimentalist School. *Journal of Economic Perspectives*, 24(2): 47-58.

Keane, M.P. and N. Wasi (2013), "Comparing Alternative Models of Heterogeneity in Consumer Choice Behavior," *Journal of Applied Econometrics*, forthcoming.

Keane, M. and R. Sauer (2010). A Computationally Practical Simulation Estimation Algorithm for Dynamic Panel Data Models with Unobserved Endogenous State Variables, International Economic Review, 51:4 (November), 925-958.

Keane, M. and K. Wolpin (2001). The Effect of Parental Transfers and Borrowing Constraints on Educational Attainment. International Economic Review, 42:4, 1051-1103.

Keller, Kevin (2002). "Branding and Brand Equity," in B. Weitz and R. Wensley (eds.), *Handbook of Marketing*, Sage Publications, London, p. 151-178

Koopmans, T.C., H. Rubin and R.B. Leipnik (1950). Measuring the Equation Systems of Dynamic Economics. Cowles Commission Monograph No. 10: Statistical Inference in Dynamic Economic Models, T.C. Koopmans (ed.), John Wiley & Sons, New York.

Lancaster, Kelvin J. (1966), "A New Approach to Consumer Theory," *Journal of Political Economy*, 74, 132-157.

Lerman, S. and C. Manski, C. (1981), 'On the use of simulated frequencies to approximate choice probabilities', in C. Manski and D. McFadden, eds., Structural Analysis of Discrete Data with Econometric Applications, MIT Press, Cambridge, MA, 305–319.

Mace, S. and S. Neslin (2004), "The Determinants of Pre- and Postpromotion Dips in Sales of Frequently Purchased Goods," *Journal of Marketing Research*, 41:3, 339-350.

McFadden, D. (1974), Conditional Logit Analysis of Qualitative Choice Behavior, in *Frontiers in Econometrics*, in P. Zarembka (ed.), New York: Academic Press, 105-42.

McFadden, D., "A Method of Simulated Moments for the Estimation of Discrete Response Models without Numerical Integration," Econometrica, 57:5 (1989), 995-1026.

McFadden, D. and K. Train (2000), "Mixed MNL models for discrete response," *Journal of Applied Econometrics,* 15, 447-470.

Neslin, Scott. (2002). Sales Promotion. Cambridge: Marketing Science Institute, Relevant Knowledge Series.

Neslin, Scott A. (2002), "Sales Promotion," in *Handbook of Marketing*, edited by Barton A. Weitz and Robin Wensley, London: Sage Publications

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

Pauwels, K., D. Hanssens and S. Siddarth (2002), "The Long-Term Effects of Price Promotions on Category Incidence, Brand Choice, and Purchase Quantity," Journal of Marketing Research, 39:4, 421-39.

Pesendorfer, Martin. (2002). ''Retail Sales: A Study of Pricing Behavior in Supermarkets'', Journal of Business 75(1), 33–66.

Rossi, P., Allenby, G. and R. McCulloch (2005), *Bayesian Statistics and Marketing*, John Wiley and Sons, Hoboken, N.J..

Smith, Martin D., 2005. "State dependence and heterogeneity in fishing location choice," *Journal of Environmental Economics and Management*, 50(2), pages 319-340, September.

Srinivasan, T.C. and Russell S. Winer. (1994). ‘‘Using Neoclassical Consumer-Choice Theory to Produce a Market Map From Purchase Data’’, Journal of Business and Economic Statistics 12 (January), 1–9.

Stern, Steven, 1992. "A Method for Smoothing Simulated Moments of Discrete Probabilities in Multinomial Probit Models," Econometrica, Econometric Society, vol. 60(4), pages 943-52, July.

Stigler, George J., 1984. "Economics—The Imperial Science?" *Scandinavian Journal of Economics*, 86(3), pp. 301-313.

Sun, B., S. Neslin and K. Srinivasan (2003), "Measuring the impact of promotions on brand switching under rational consumer behavior," *Journal of Marketing Research*, 40:4, 389-405.

Train, K. (2003), *Discrete Choice Methods with Simulation*, Cambridge University Press.

Van Heerde, S. Gupta and D. Wittink (2003). "Is 75% of the sales promotion bump due to brand switching? No, only 33% is," *Journal of Marketing Research*, 40:4, 481-491.

Van Heerde, H., P. Leeflang and D. Wittink (2000), "The Estimation of Pre- and Postpromotion Dips with Store-Level Scanner Data," *Journal of Marketing Research*, 383-95.

Van Heerde, H., P. Leeflang and D. Wittink (2004), "Decomposing the sales promotion bump with store data," *Marketing Science*, 23:3, 317-334.

Winer, Russell S. (1986). ‘‘A Reference Price Model of Brand Choice for Frequently Purchased Products’’, Journal of Consumer Research 13(September), 250–256.

Wooldridge J. 2003a. *Econometric Analysis of Cross Section and Panel Data*. MIT Press: Cambridge, MA.

Wooldridge J. 2003b. Simple Solutions to the Initial Conditions Problem is Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity. Working Paper, Michigan State University.