The Value of Purchase History Data in Target Marketing
Author(s): Peter E. Rossi, Robert E. McCulloch and Greg M. Allenby
Source: *Marketing Science,* Vol. 15, No. 4 (1996), pp. 321-340
Published by: INFORMS
Stable URL: http://www.jstor.org/stable/184168
Accessed: 21-01-2018 13:18 UTC

REFERENCES
Linked references are available on JSTOR for this article:
http://www.jstor.org/stable/184168?seq=1&cid=pdf-reference#references_tab_contents
You may need to log in to JSTOR to access the linked references.

The Value of Purchase History Data in Target Marketing

Peter E. Rossi • Robert E. McCulloch • Greg M. Allenby

University of Chicago
University of Chicago
Ohio State University

Abstract

An important aspect of marketing practice is the targeting of consumer segments for differential promotional activity. The premise of this activity is that there exist distinct segments of homogeneous consumers who can be identified by readily available demographic information. The increased availability of individual consumer panel data open the possibility of direct targeting of individual households. The goal of this paper is to assess the information content of various information sets available for direct marketing purposes. Information on the consumer is obtained from the current and past purchase history as well as demographic characteristics. We consider the situation in which the marketer may have access to a reasonably long purchase history which includes both the products purchased and information on the causal environment. Short of this complete purchase history, we also consider more limited information sets which consist of only the current purchase occasion or only information on past product choice without causal variables.

Proper evaluation of this information requires a flexible model of heterogeneity which can accommodate observable and unobservable heterogeneity as well as produce household level inferences for targeting purposes. We develop new econometric methods to implement a random coefficient choice model in which the heterogeneity distribution is related to observable demographics. We couple this approach to modeling heterogeneity with a target couponing problem in which coupons are customized to specific households on the basis of various information sets. The couponing problem allows us to place a monetary value on the information sets.

Our results indicate there exists a tremendous potential for improving the profitability of direct marketing efforts by more fully utilizing household purchase histories. Even rather short purchase histories can produce a net gain in revenue from target couponing which is 2.5 times the gain from blanket couponing. The most popular current electronic couponing trigger strategy uses only one observation to customize the delivery of coupons. Surprisingly, even the information contained in observing one purchase occasion boasts net couponing revenue by 50% more than that which would be gained by the blanket strategy. This result, coupled with increased competitive pressures, will force targeted marketing strategies to become much more prevalent in the future than they are today.

(*Target Marketing; Coupons; Heterogeneity; Bayesian Hierarchical Models*)

1. Introduction

In consumer marketing, it has long been recognized that it would be potentially valuable to customize marketing activities. Unfortunately, the lack of accurate information about specific household preferences has made it difficult to price and promote products at the individual household level. For example, most coupons are not customized to specific households and are mass-distributed via newspaper inserts. Recently, electronic distribution of coupons has become more widespread under programs such as Catalina Marketing Incorporated's (CMI) Checkout Coupon and various Frequent Shopper schemes in which household receive volume discounts. These electronic programs hold out the possibility of customizing the coupon to specific households. For example, CMI's program triggers coupons based on current shopping purchases. More complete purchase history data is often available in Frequent Shopper programs. The most ambitious program to date is the CMI Checkout Direct couponing service which can create actual panel history data on customers including both purchase history and information on the store causal environment. Most direct marketers already collect extensive household information but rarely exploit this information to customize the product offerings or merchandizing strategies. Instead, direct marketers have focused on the scoring of households for receipt of fairly uniform marketing efforts.

The continued decline in information processing and storage costs will make the collection of purchase history information even more cost effective in the future. The real question is whether or not it will be worthwhile for marketers to collect and exploit more detailed and complete purchase history information. For example, it may be possible to acquire, at low cost, demographic information about a specific household. The marketer may then consider adding some purchase history data which may be very short (one observation) or many observations and which may include or not include causal information about alternative products. The goal of this paper is to value these expanding information sets. To provide a metric for valuation, we consider a target couponing problem in which we allow for the possibility of customizing the face value of the coupon to specific households.

Proper valuation of household information requires a flexible statistical model which can incorporate both observable and unobservable heterogeneity. Because customization requires inference about individual household parameters on the basis of as little as one observation, we require that the inference framework not rely on large sample approximations. The market acceptance of the CMI checkout coupon program suggests that there may be valuable information even in a single observation.

Our results indicate there exists a tremendous potential for improving the profitability of direct marketing efforts by more fully utilizing household purchase histories. Even rather short purchase histories can produce a net gain in revenue from target couponing which is 2.5 times the gain from blanket couponing. The most popular current electronic couponing trigger strategy uses only one observation to customize the delivery of coupons. Surprisingly, even the information contained in observing one purchase occasion boasts net couponing revenue by 50% more than that which would be gained by the blanket strategy.

The remainder of the paper is organized as follows: Section 2 presents our random effects choice model used to measure household preferences and sensitivities. Section 3 presents our full Bayesian method of inference for this model. Section 4 lays out the various information sets which will be used to infer about household level parameters. Section 5 discusses the data and provides inferences about the common parameters. Section 6 illustrates our methods for inferring about households parameters conditional on various information sets. A target couponing problem is proposed in §7 to provide a metric for assessing information content. Section 8 presents our results on the value of various information sets in the target couponing exercise. Concluding remarks are offered in §9.

2. A Flexible Random Effects Model for Consumer Heterogeneity

To assess the information content of purchase history data, we start by postulating a flexible model of brand choice. We observe the multinomial choice outcome, I_{ht}, for household h at time t conditional on a set of explanatory covariates. In a standard random utility frame-

work (McFadden 1974), the choice outcome stems from an m-dimensional vector of latent utilities which follows a multivariate regression

$$y_{h,t} = X_{h,t}\beta_h + \epsilon_{h,t}, \quad \epsilon_{h,t} \sim N(0, \Lambda),$$

$$h = 1, \ldots, H, \quad t = 1, \ldots, T_h. \tag{1}$$

Here there are m brand choices whose utilities are given by $y_{h,t}$. $X_{h,t}$ is a matrix of choice characteristics which includes an intercept term for each of the m brands and price, feature and display variables, $X_{h,t} = [D_m, p, d, f]$ where D_m a diagonal matrix with ones in every diagonal position except the first which is zero (for identification purposes), p is an m vector of log prices, d is an indicator vector of length m such that the ith element is 1 if the ith brand is on display and 0 otherwise, and f is an indicator vector for feature. β_h is a vector representing the household h's preferences and sensitivity to marketing mix variables and $\epsilon_{h,t}$ is an error term. Household h has a purchase history of length T_h. We do not observe the latent utilities (y_{ht}) directly but simply the choice which indexes of the maximum (I_{ht}, $I_{ht} \in \{1, 2, \ldots, m\}$). Thus, the sampling model consists of the censoring mechanism which produces the multinomial outcome and the latent multivariate regression

$$I_{ht} \mid y_{ht}, \tag{2a}$$

$$y_{ht} \mid X_{ht}, \beta_h, \Lambda. \tag{2b}$$

Throughout the paper, the notation "$y \mid x$" means conditional distribution of the random variable y given x.

Different specifications of the error structure result in various probit (normal error) and logit (extreme value error) models. We use a diagonal covariance structure $\epsilon_{h,t} \sim$ iid $N(0, \Lambda)$, where Λ is a $m \times m$ diagonal matrix, coupled with the identifying assumption that the first diagonal element is one. The diagonal covariance structure greatly simplifies the calculation of the choice probabilities needed in the subsequent analysis while avoiding the restrictive IIA property associated with a scalar covariance structure (Allenby and Ginter 1995). Unobserved heterogeneity, as discussed below, serves to introduce a correlated error structure as in Hausman and Wise (1978).

To model heterogeneity across households, we adopt a flexible random coefficient model with heterogeneity driven by both observable (demographic) characteris-

tics of the household and unobservable components. Our approach is to model the mean of the coefficient vector (β_h) as a function of demographic variables. That is, we allow both the slopes and intercepts to vary around a systematic component which is determined in a multivariate regression of $\{\beta_h\}$ on a set of d demographic variables.

$$\beta_h = \Delta z_h + v_h, \quad v_h \sim \text{iid } N(0, V_\beta), \quad h = 1, \ldots, H. \tag{3}$$

z_h is $d \times 1$ vector consisting of an intercept and $d - 1$ demographic variables. Δ is a $k \times d$ matrix of regression coefficients where k is the number of brands plus the number of causal variables. This specification allows the preferences or intercepts to vary by demographic variables as well as the slopes. The multivariate regression introduces $d*k$ new parameters to accommodate different regression relationships for each of the k β_h coefficients. The magnitude of V_β determines the dispersion of the distribution of unobserved heterogeneity. By comparing the marginal variance of β_h with V_β, we can assess how much of the variability of β_h can be ascribed to observable vs. unobservable heterogeneity.

An alternative approach to modelling consumer heterogeneity is a finite mixture model of the sort pioneered by Kamakura and Russell (1989) in which the random coefficient model in (3) is replaced by a discrete distribution. That is, the $\{\beta_h\}$ vectors are drawn from a discrete distribution with multiple mass points and the Λ parameters remain fixed across households. In Appendix B we compare the continuous and finite mixture approaches. While the finite mixture model of heterogeneity can capture many important features of heterogeneity, it appears to us that, in this application to estimating household level parameters, the continuous mixture model provides a better fit to the data and more reasonable inferences.

Our approach to modelling heterogeneity in the probit model builds up the model specification through a series of conditional distributions in what is commonly called a hierarchical model:

$$I_{ht} \mid y_{ht}, \tag{4a}$$

$$y_{ht} \mid X_{ht}, \beta_h, \Lambda, \tag{4b}$$

$$\beta_h \mid z_h, \Delta, V_\beta. \tag{4c}$$

We adopt a Bayesian approach to conducting inference in this hierarchical probit model rather than a standard classical econometric approach (c.f. Borsch-Supan and Hajivassiliou (1992), Hajivassiliou (1993), or Keane (1994)) for two reasons: (1) customized marketing actions regarding individual households require inferences about the household parameters $\{\beta_h\}$ directly and not just the common parameters (Δ, V_β, Λ), and (2) since we are making inferences in many cases on the basis of only a handful of observations we need a method which properly accounts for parameter uncertainty and is free from approximations which rely on large sample asymptotic theory.

To complete the model, we introduce priors over the parameters which are common to all households, $\{\Lambda, \Delta, V_\beta\}$. We specify a prior on Δ and V_β in the natural conjugate form where Δ is normal given V_β with hyperparameters $\bar{\Delta}$ and A_d and V_β is inverted Wishart with hyperparameters v_b and V_{ob}. The prior on V_β and the sample information are used to infer about the relative diffusion of the household coefficients. Finally, we complete our hierarchical model with a prior over the variance of the random utility errors, Λ, in the form of independent inverted gamma distributions (see Appendix A for details).

Both the setting of prior hyperparameters and the nature of the prior distribution have the potential to influence the posterior distribution of the $\{\beta_h\}$. In practice, we take very diffuse priors over Δ and Λ and induce a mild amount of shrinkage with our V_β prior (see Appendix A for the exact parameter settings). To investigate alternative forms of the prior distribution, we consider a finite mixture random effects model in Appendix B.

3. Posterior Computations

Bayesian analysis of hierarchical models has been made feasible by the development of Markov chain simulation methods which directly exploit the hierarchical structure. (See Tanner and Wong 1987, Gelfand and Smith 1990, Gelfand et al. 1991, and Tierney 1991 for general discussion of these methods.) The basic idea behind these methods is to construct a Markov chain which has the posterior as its stationary or invariant distribution and then simulate the chain to obtain a se-

quence of draws which can be used to approximate the posterior to any desired degree of accuracy. In this paper, we use the Gibbs sampler constructed for the hierarchical MNP model by McCulloch and Rossi (1994). The Gibbs sampler is implemented by drawing successively from the following set of posterior distributions which are based on the data consisting of the X_{ht} explanatory variables and $I_{h,t}$ (the index of observed choices):

$$y_{h,t} \mid I_{h,t}, \beta_h, \Lambda, X_{h,t}, \quad h = 1, \ldots, H, \ t = 1, \ldots, T_h, \quad (5a)$$

$$\beta_h \mid y_{h,t}, \Lambda, \Delta, V_\beta, X_{h,t}, z_h, \quad h = 1, \ldots, H, \quad (5b)$$

$$\Lambda \mid \{y_{h,t}\}, \{\beta_h\}, \{X_{h,t}\}, \quad (5c)$$

$$\Delta \mid \{\beta_h\}, V_\beta, \{z_h\}, \quad (5d)$$

$$V_\beta \mid \{\beta_h\}, \Delta, \{z_h\}. \quad (5e)$$

The exact forms for these conditional distributions are given in Appendix A.

Our Gibbs sampler proceeds by drawing successively from each of the distributions above and iterating this procedure to obtain a long sequence of draws. These draws are then used to compute the marginal posterior distribution of various quantities of interest. There are a number of technical issues which arise in using these sorts of procedures. (See McCulloch and Rossi 1994 and Gelman and Rubin 1992 for a thorough discussion of these issues.)

The ultimate objective of this paper is to evaluate the benefits from various target marketing activities which can be developed from having access to some sort of household level information. We can think of all market segmentation and customized marketing activities as based on some (usually partial) information set. For example, market segmentation based on observed demographics uses only the demographic information about a household and not purchase history information. Even with complete purchase history data, we will not be able to infer about the household level parameters with very high precision. It is, therefore, imperative to develop methods which characterize the uncertainty about these household parameters. Our Bayesian methods are ideal in this regard since we obtain the entire posterior distribution for each household parameter as a by-product of the Gibbs sampler. In the sections below, we will experiment with posteriors based on dif-

324

ferent information sets to establish the incremental value of various sorts of household level information.

4. Alternative Information Sets

In order to begin an evaluation of the worth of household purchase history information, we must delineate the information sets upon which different target marketing actions can be based. All targeting strategies are based on inferences about a specific household's preferences. If we were able to infer with perfect certainty about a household's preferences, we could customize (subject to transactions costs) many aspects of merchandizing including pricing and couponing. However, we rarely have access to sufficient household data to make very accurate inferences, and the optimal customization procedure must strike the right trade-off between the incentive to customize which comes from variation in preferences across households and our limited ability to infer about preferences.

Base Information Set

In many consumer marketing situations, we have no specific information about our potential customer's preferences. Our only information is about the distribution of preferences across customers. In these situations, only one merchandizing strategy can be taken for all consumers in the market. For example, a blanket coupon drop is designed on the basis of the distribution of brand preferences and price sensitivity in the population. Firms acquire information on the distribution of preferences from a variety of forms of consumer research, including information on buying habits from panels of households maintained by Nielsen or IRI. These panels can be used to assess the distribution of brand preferences and sensitivities to marketing mix variables. However, these panels cannot be used for targeting purposes since the panel members are a small fraction of the market.

To approximate the base information set available to many firms, we use the predictive distribution of our probit choice model parameters, β_h, based on the posterior distribution of the heterogeneity parameters, Δ and V_β. That is, we think of our panel sample as giving us information on the distribution of preferences across households through our set of purchase history records. A useful way to view this process is to think of inferring

about $\beta_{h'}$ for household h' which is a random draw from the distribution of all households. The predictive distribution represents our beliefs about $\beta_{h'}$ given the model and the data used in our panel of households. If we knew $z_{h'}$, Δ, and V_β, our model of heterogeneity would give the predictive distribution of $\beta_{h'}$ as $N(\Delta z_{h'}, V_\beta)$. Of course, we don't know $z_{h'}$ so that we must integrate out $z_{h'}$ using some distribution of $z_{h'}$ which represents our beliefs about the likely values

$$p(\beta_{h'} \mid \Delta, V_\beta) = \int p(\beta_{h'} \mid z_{h'}, \Delta, V_\beta) p(z_{h'}) dz_{h'}. \quad (6)$$

In all of what follows, we simply use the empirical distribution of z_h in our panel dataset. That is, we average the normal distribution of $\beta_{h'}$ over all of the z_h vectors in the dataset. It should be noted that this empirical distribution could easily be replaced by any distribution of demographics. For example, if our method were to be applied to a different population of households, the empirical distribution of demographics for this new target population could be substituted for the empirical distribution in our sample of households.

It is also the case that our information regarding the common Δ and V_β parameters is not perfect. The uncertainty in our knowledge of these parameters is reflected in the posterior distribution from the sample of H households. The final predictive distribution is produced by integrating or averaging the predictive distribution conditional on Δ and V_β with respect to the posterior.

$$p(\beta_{h'} \mid \text{Data}) = \int p(\beta_{h'} \mid \Delta, V_\beta) p(\Delta, V_\beta \mid \text{Data}) d\Delta dV_\beta. \quad (7)$$

To compute the integrals in the above equation, we draw $z_{h'}$ from the empirical distribution of the demographics and use the draws for (Δ, V_β) which are produced as a natural by-product of the Gibbs sampler. Conditional on each of the draws, we then sample the corresponding normal distribution. Thus, while the base information set predictive distribution is of no known analytic form, we can easily define a simulator which can approximate it to any desired degree of accuracy.

Demographic Information Set

The next information set expands from the base set to include demographic information about specific consumers. Many market segmentation schemes are based on dividing the population of consumers into distinct groups on the basis of observation demographic attributes. We append demographic information to the base information set of the distribution of consumer preferences. Given the base of information on the universe of households, we observe demographic variables for a new household and condition our inferences about β_h on that information. For the new household indexed by h', we observe $z_{h'}$ the vector of demographic information and we must calculate the posterior distribution of $\beta_{h'}$ given $z_{h'}$ and the data in the base information set:

$$p(\beta_{h'} \mid z_{h'}, \text{Data})$$

$$= \int p(\beta_{h'} \mid z_{h'}, V_\beta, \Delta) p(V_\beta, \Delta \mid \text{Data}) dV_\beta d\Delta. \quad (8)$$

The only difference between the base and demographic information sets is that to compute the distribution of β_h for the demographic information set we condition on the observed demographics of the household. Thus, we don't require the further step of integrating or averaging over the empirical distribution of z_h.

Choices-only Information Set

Many frequent shopper and buyer loyalty programs collect data on the actual choices of a large set of households. In most cases, these programs do not collect information on the causal environment confronting the consumer. In many instances, these frequent shopper programs enroll more than 25% of store shoppers; examples include Dominick's Finer Foods in Chicago and UKrops in Virginia. This data is currently used only to make periodic store loyalty rewards which include discounts on future purchases and large manufacturer coupons. However, this data could also be used to infer about the preferences of particular households for the purposes of designing custom coupon programs. A more direct example of customization is the Catalina Marketing Inc.'s "Checkout Direct" product which uses check cashing and credit card numbers to piece together purchase history records for specific consumers. This enables Catalina to implement coupon trigger strategies which depend on the whole purchase history. Cur-

rently, the "Checkout Direct" product does not retain information on the causal environment facing the consumer, although this is technologically feasible.

In our case, the causal variables which might not be observed are price, display, and feature. Recall our model for the latent utilities from (1):

$$y_{ht} = X_{ht}\beta_h + \epsilon_{ht} = \begin{bmatrix} 0 \\ \gamma_h \end{bmatrix} + X_{c,ht}\delta_h + \epsilon_{ht},$$

$$\epsilon_{ht} \sim N(0, \Lambda). \quad (9)$$

Here X_{cht} is the matrix of values of causal variables, $[p, d, f]$. $\gamma_h' = [\gamma_{h,2}, \ldots, \gamma_{h,p}]$. In order to implement the target couponing problem, we must be able to make inferences about $\beta_h' = (\gamma_h', \delta_h')$ for households for which we do not observe X_c. For these households, we have no choice but to fit a model which does not include X_c as defined below

$$y_{ht} = \begin{bmatrix} 0 \\ \mu_h \end{bmatrix} + \epsilon_{ht}, \quad \epsilon_{ht} \sim N(0, \Lambda'). \quad (10)$$

$\mu_h' = [\mu_{h,2}, \ldots, \mu_{h,p}]$. The question, then, is how to convert information about μ_h into information about complete parameter vector, β_h.

It is clear that we must bring to bear other sources of information to successfully solve this problem. We use information on households for which all variables are observed to help us infer the complete parameter vector for households lacking causal information. For example, consider the case in which the only missing variable is price. We observe a household purchasing brand A on most choice occasions, but no price information is available for this household. We look at households with similar buying patterns (namely, loyalty to brand A) for which we have complete information and impute a similar price sensitivity.

More formally, we must map the posterior distribution of μ_h into a distribution on $\beta_h' = (\gamma_h', \delta_h')$. (To reduce notational clutter, we drop the h subscript on all that follows in this section.) We must bear in mind that the parameters in the model with no causal variables have a different interpretation than in the full model. There are two reasons for this. First, γ, the vector of intercepts in a multivariate regression with covariates, is not the same as the unconditional mean of y given by μ. Second, the identification restrictions are applied differently in

the two models. Starting with the full model in (9), we derive the reduced model to establish the correspondence between γ and μ. Taking iterated expectations, we see $E[y] = E_x[E[y \mid X]] = E_x[X\beta]$ or

$$E[y] = \begin{bmatrix} \bar{x}_1'\delta \\ \gamma_2 + \bar{x}_2'\delta \\ \vdots \\ \gamma_p + \bar{x}_p'\delta \end{bmatrix}. \tag{11}$$

$\bar{x}_j$ is the vector of means of the causal variables for the jth alternative (e.g., average price, display and feature for the jth brand). This is where prior information on the distribution of causal variables is required.

We cannot directly use Equation (11) above to relate μ to γ since a different identification restriction is imposed on each. To identify the parameters in the model estimated with no intercepts, we set the first intercept to zero as in Equation (10) above. To insure an equivalence in the two parameterizations, we must set the first element of unconditional mean derived in (11) to zero by subtracting the first element from the vector. This gives us a relationship between μ and γ, δ.

$$\mu_2 = \gamma_2 + \bar{x}_2' - \bar{x}_1', \ldots, \mu_p = \gamma_p + \bar{x}_p' - \bar{x}_1'$$

or, in matrix form,

$$\gamma + R\delta \equiv \mu \quad \text{where } R = \begin{bmatrix} \bar{x}_2' - \bar{x}_1' \\ \bar{x}_3' - \bar{x}_1' \\ \vdots \\ \bar{x}_p' - \bar{x}_1' \end{bmatrix}. \tag{12}$$

Our goal, then, is to make inferences about the β vector given information in the posterior distribution of μ. If we knew μ, we could compute the conditional distribution of $\delta \mid \mu$ from our model of heterogeneity since (12) expresses δ as a linear function of μ and γ:

$$\begin{bmatrix} \gamma \\ \delta \end{bmatrix} \sim N(\Delta d, V_\beta), \tag{13}$$

$$\begin{bmatrix} \mu \\ \delta \end{bmatrix} = \begin{bmatrix} I_{p-1} & R \\ 0 & I_{k-p+1} \end{bmatrix} \begin{bmatrix} \gamma \\ \delta \end{bmatrix} = A \begin{bmatrix} \gamma \\ \delta \end{bmatrix}. \tag{14}$$

Therefore,

$$\begin{bmatrix} \mu \\ \delta \end{bmatrix} \sim N(A\Delta d, AV_\beta A') \equiv N(\tau, \Sigma). \tag{15}$$

We then can compute the distribution of $\delta \mid \mu, \Delta, V_\beta$ from standard conditional normal theory:

$$\delta \mid \mu \sim N(E[\delta \mid \mu], \text{Var}(\delta \mid \mu)),$$

$$E[\delta \mid \mu] = \tau_\delta + \Sigma_{\delta,\mu}\Sigma_{\mu,\mu}^{-1}(\mu - \tau_\mu),$$

$$\text{Var}(\delta \mid \mu) = \Sigma_{\delta,\delta} - \Sigma_{\delta,\mu}\Sigma_{\mu,\mu}^{-1}\Sigma_{\mu,\delta}. \tag{16}$$

We can uncondition on μ, Δ, V_β by using the draws from posterior distributions of these quantities.

We implement our algorithm for making inferences on β given the information in the purchase history as follows:

(i) Compute the posterior distribution of μ by running our standard Gibbs sampler on dataset with no causal variables

(ii) Choose a draw of Δ, V_β from our analysis of a full information subset of households

(iii) Choose a draw of μ from the analysis in (i)

(iv) Draw $\delta \mid \mu$, Δ, V_β, d from (16) above

(v) Compute the value of γ corresponding to δ from Equation (14)

(vi) repeat for all draws of (μ, Δ, V_β).

Full Information Set

The most complete information set would consist of purchase histories for individual consumers, including information on the causal environment in the store. Although this sort of information is rarely collected or exploited by retailers and manufacturers today, there is no question that this is currently technologically feasible. For example, store scanner data and audits for displays and features could be combined with standard frequent shopper program data to build up both the consumer purchase and causal environment variables. We expect the costs of retaining and processing purchase history data will continue to decline, which will force retailers and manufacturers to consider the value of full purchase history information.

Even in a world in which retailers are committed to retaining purchase history and causal information on a substantial subset of consumers, it is unlikely that very long histories will be obtained. Consumers are mobile, and data storage capacity is limited. An extreme example of this is the Catalina Marketing Inc.'s "Checkout Coupon" electronic couponing service. In this product, coupons are triggered by information obtained at the point of purchase for only *one* purchase occasion. The Catalina coupon computer/printer interfaces directly with the P-O-S checkout register and can look up the

Table 1   Information Sets

Name	Description
Base	Information on the distribution of preferences in the population—no specific information about individual consumers
Demographic	Base Info + demographic information on each specific consumer
Choices-only	Demographic Set + information on purchase history for specific consumers—no information on causal environment
One Observation	Demographic Set + Brand Choice and causal information are available for one purchase occasion. This information set simulates the information available in current electronic couponing schemes
Full	Complete information on purchase history and causal environment

prices of competitive products from the store computer. While the majority of trigger strategies used by Catalina today use only choice information, some current Catalina programs do use competitive price information. Catalina estimates that their electronic couponing machines are installed in stores which represent over 25% of the ACV in grocery today. The success of the Catalina product suggests that there is valuable information even in only datasets which consist of only one observation (including causal variables) per household.

With access to the full information set, we can infer about β_h based on our knowledge of the marginal distribution of β_h in the population, the specific value of the demographic attributes, and some history (although possibly very limited) of choice behavior for that household. To update our inferences to this full information set, we must simply compute the posterior distribution of $\beta_{h'}$ given $z_{h'}$, $(I_{h'1}, \ldots, I_{h'T_{h'}})$, $(X_{h'1}, \ldots, X_{h'T_{h'}})$ and the information in the base population of households. This is given automatically by the posterior distributions constructed from the Gibbs sampler run with the full sample of households. To evaluate the "Catalina" or "One Observation" dataset value, we simply delete all but one observation for a random subset of households in our dataset. The "One Observation" data set includes both choice and causal information.

A summary of all the information sets is given in Table 1.

5.   Data and Inferences About Common Parameters

The data used in our analysis is an A. C. Nielsen scanner panel dataset of tuna purchases in Springfield, Missouri. Five brands of tuna packaged in six-ounce cans, which account for 75% of the total category volume, are included in the analysis. 400 households are selected at random from the 775 households who remained in the panel at least 1.5 years. These households make, on average, 13 purchases from this set of five brands with a range of between 1 and 61 purchases. Price is entered into the model in logarithmic form. In addition, the existence of in-store displays and feature advertisements at the time of purchase are represented by dummy variables. A summary of the brands, their average price and level of merchandising support are provided in Table 2. To identify the model, we set the intercept for the Chicken of the Sea (C-O-S) water-packed tuna to zero; thus, all results must be interpreted as relative to the C-O-S water brand.

Six demographic variables (z_h) are included in the analysis: household income, family size, a retirement indicator variable equal to one if the head of household is retired, an indicator of whether the head of household is unemployed and an indicator of female headed families. Table 3 provides summary statistics for these variables. These variables were selected on the grounds of economic and marketing plausibility, as well as the fact that they exhibit substantial variability across households.

The relative value of demographic vs. purchase history information will hinge on how much of the vari-

Table 2   Description of the Data

Alternative	Choice Share	Average Price	% of Time Displayed	% of Time Featured
Packaged in Water:				
Chicken of the Sea	0.413	0.681	0.201	0.372
Starkist	0.294	0.758	0.126	0.256
House Brand	0.053	0.636	0.101	0.142
Packaged in oil:				
Chicken of the Sea	0.134	0.694	0.164	0.307
Starkist	0.104	0.751	0.195	0.232

Table 3 Household Demographic Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
Household Income	27,739	15,320	2,500	130,000
Family Size	3.159	1.340	1.0	9.0
Retired	0.165	0.372	0.0	1.0
Unemployed*	0.06	0.24	0.0	1.0
Wsmom†	0.08	0.27	0.0	1.0

* Labor force status of male head of household (or female head if male not present). Defined as unemployed if employed part-time, temporarily unemployed, or disabled.

† Family size ≥ 2 and male head of household not present.

ability of the household specific parameters can be explained by observable demographic characteristics as opposed to unobserved heterogeneity. Our model of heterogeneity provides a natural way of assessing this by examination of the posterior distribution of the Δ coefficient matrix and the unobserved heterogeneity or diagonal terms of the V_β matrix. The hierarchical model introduced above is multivariate regression in which

each of the k elements of β_h is regressed on d demographic variables:

$$\beta_h = \Delta z_h + v_h \quad \text{with } v_h \sim N(0, V_\beta). \quad (17)$$

Table 4 presents information on the posterior distribution of Δ. The table presents the Bayes estimates (posterior means) along with the posterior probability that the coefficient is negative or positive (depending on the sign of the estimate). In addition, the posterior means of the square root of the diagonal elements of V_β are given in the last column. The demographic variables have been coded in terms of deviation from the variable mean so that the Δ elements corresponding to the constant column (denoted "Cons" in the table) are the expected value of β_h for average demographic values. Rather than sorting through variables in a specification search, our approach is to leave demographic variables in each of the equations even if their posterior distributions put a great deal of mass near zero.

It is interesting to consider some of the regressions in the rows of Table 4. The row corresponding to the Private Label intercept shows that unemployed and

Table 4 Posterior Distribution of Delta Coefficients

Beta	Cons	ln(Inc)	ln(Fam Size)	Retire	Unemp HH	Single Mom	Unobs. Hetero.†	ρ^2 *
Starkist Water Int	**-0.11** [0.95]	0.15 (0.88)	-0.022 [0.55]	-0.05 [0.58]	**0.74** (0.99)	-0.23 [0.80]	1.01	0.20
Private Label Int	**-4.3** [1.0]	**-1.2** [1.0]	0.35 (0.76)	-0.23 [0.65]	**2.47** (1.0)	-0.44 [0.75]	2.49	0.33
C-O-S Oil Int	**-1.9** [1.0]	-0.26 [0.88]	0.21 (0.76)	0.31 (0.78)	0.55 (0.85)	0.067 (0.55)	2.17	0.13
Starkist Oil Int	**-2.4** [1.0]	-0.19 [0.75]	-0.14 [0.64]	0.43 (0.81)	**1.7** (0.99)	-0.059 [0.54]	2.62	0.095
Price Coef	**-7.3** [1.0]	-0.26 [0.73]	-0.56 [0.83]	**-1.7** [0.98]	-0.39 [0.67]	-0.16 [0.54]	3.31	0.073
Display Coeff	**0.26** [1.0]	-0.052 (0.64)	-0.039 [0.58]	**-0.43** [0.96]	**-0.58** [0.96]	0.078 [0.54]	0.64	0.132
Feature Coef	**0.41** (1.0)	0.038 (0.63)	0.14 (0.82)	**0.44** (0.98)	**0.42** (0.93)	0.094 (0.65)	0.65	0.120

† Unobservable heterogeneity as measured by the posterior mean of the square root of the diagonal elements of V_β.
* $\rho^2 \equiv 1 - \text{Var(unobservable component)}/\text{Var}(\beta)$.
( ) indicates probability that the coefficient is positive.
[ ] indicates probability that the coefficient is negative.
bold indicates probability exceeds 0.90.

lower income households tend to favor the Private Label, even controlling for the price differential between the Private Label and national brands. It is interesting that price sensitivity does not seem to be related in any strong way to the demographic variables as indicated by the low R^2. On the other hand, retired and unemployed people, who presumably have a lot of time on their hands, respond more to feature advertisements in the newspaper.

The column in Table 4 entitled "Unobs. Hetero." reports the posterior mean of the square root of the diagonal elements of V_β. V_β is the covariance matrix of the distribution of β_h across households after conditioning on the observable demographic characteristics. The diagonal elements of V_β measure the dispersion of household parameters across households. The figures in Table 4 show that there is tremendous variation around the mean function of the demographic variables. For example, price coefficients average around -7.0 with a standard deviation of 3.0.

The large size of the unobservable heterogeneity suggests that the demographic information may have limited value in predicting many of the key parameters. To measure this more formally, we can compute an R^2-like quantity for each of the regression coefficients in β_h. The diagonal elements of V_β measure the conditional or "error" variance of β_h; this can be combined with the marginal or unconditional variability of the elements of β_h to produce an estimate of the population coefficient of determination which is labelled "ρ^2" in the right-most column of Table 4. $\rho^2 = 1 - \text{Var}(\epsilon)/\text{Var}(\beta)$, where $\text{Var}(\epsilon)$ is the variance of the unobservable component and $\text{Var}(\beta)$ is the total variation. The demographic variables explain between 7 and 33% of the household variation in parameter values. Moreover, the demographic variables explain only 7% of the variation in price sensitivity, which is a key determinant of potential profitability for target couponing. Bucklin and Gupta (1992) and Gupta and Chintagunta (1994) find that demographics can explain even less of the variability in transformed segment probabilities.

In the next section, we formally address this question of how much information is available in purchase history data above the value of demographic information by looking at the predictive distribution of β_h conditional on various information sets.
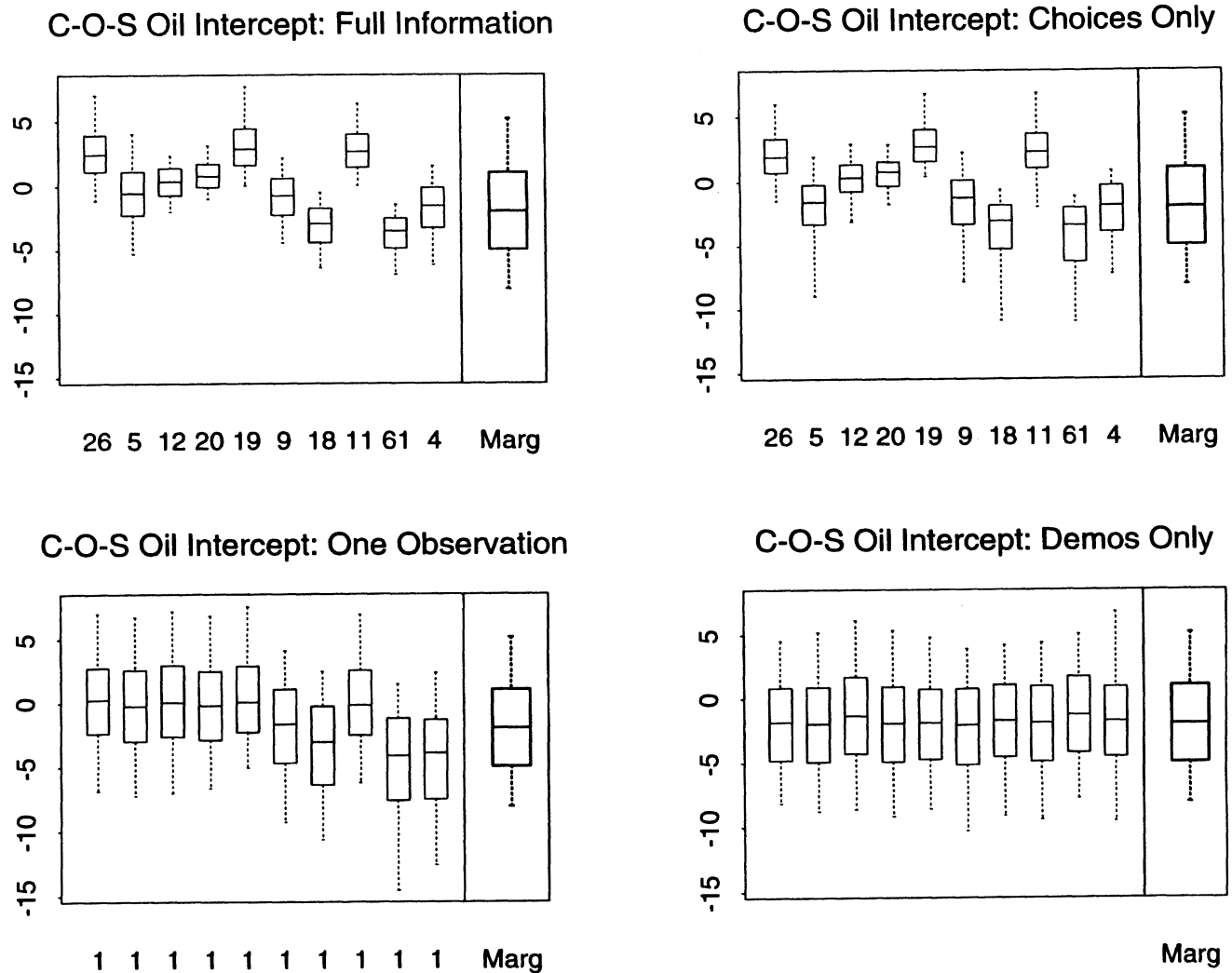
6. Alternative Information Sets and Predictive Distributions

As outlined in §2, we can construct Gibbs Samplers which will allow us to infer about household-level parameters for each of the above information sets. Figures 1 and 2 present the marginal distribution of elements of the β_h parameter for 10 of the 400 households in the our sample. The marginal distributions are displayed in a nonstandard form in which the height of the box is from the 10 to 90th percentiles (this form is used throughout the paper). It is important to note that our procedure displays the entire posterior distribution rather than just a point estimate. Given that we expect substantial uncertainty in inferring about household parameters with only a handful of observations, it is vital that we measure appropriately the uncertainty of our estimates.

Figure 1 shows the marginal posterior distribution of the C-O-S Oil brand intercept for our five information sets (see Table 1 for a summary). The top left panel shows the marginal posterior based on the full information set which includes demographics, purchase history and causal variables. Along the horizontal axis, we have provided the number of purchase occasions for each household. The boxplot labelled "Marg" is the predictive distribution of this parameter from the base information set. We can see immediately that there is a great deal of household variation in the C-O-S Oil intercept parameter. As expected, some households show a marked preference for either oil or water-packed tuna. For other households, it is hard to establish a clear preference. This can arise either because we have insufficient information to see a clear preference or because the households are purchasing both oil and water-packed brands (a portfolio effect from intra-household heterogeneity). It is our contention that models which force a classification of households into oil or water loyal groups would give misleading estimates of household behavior.

In the lower left corner, the boxplots are displayed for the same 10 households where the information set is the "One Observation" set. We simulate the One Observation information set by deleting all but the first purchase occasion for 200 of the 400 households and retaining 200 households to form the "base information" set. A dra-

Figure 1   Boxplots of posterior distributions of household intercept parameters. Various information sets. 10 selected households with the number of purchase occasions indicated along the *X* axis below each boxplot. The boxplot labelled "Marg" is the predictive distribution for a representative household from the model heterogeneity distribution. Note that these are the 11–20th households as ordered in our dataset.

C-O-S Oil Intercept: Full Information

26 5 12 20 19 9 18 11 61 4   Marg

C-O-S Oil Intercept: Choices Only

26 5 12 20 19 9 18 11 61 4   Marg

C-O-S Oil Intercept: One Observation

1 1 1 1 1 1 1 1 1 1   Marg

C-O-S Oil Intercept: Demos Only

Marg

matic reduction in information content is observed. The boxplots show much less variation in central tendency and a marked increase in variation.

In the upper right corner, the household posteriors are displayed for the "Choice-Only" information set. This information set is simulated by deleting causal information for the first 200 households. Here we see that a remarkable amount of information regarding the intercept is revealed by choice-only data. The major dif-

ference between inferences based on the full information and "choices-only" data is that the dispersion of the household distributions increases when no causal data is available as might be expected. Finally, the right bottom graph shows the marginal posteriors for the "Demographics Only" information set. As can be clearly discerned, the demographic variables provide little exploitable information regarding variation in the intercept parameter.

Figure 2 Boxplots of posterior distributions of household price coefficients. Various information sets. 10 selected households with the number of purchase occasions indicated along the *X* axis below each boxplot. The boxplot labelled "Marg" is the predictive distribution for a representative household from the model heterogeneity distribution. Note that these are the 11–20th households as ordered in our dataset.
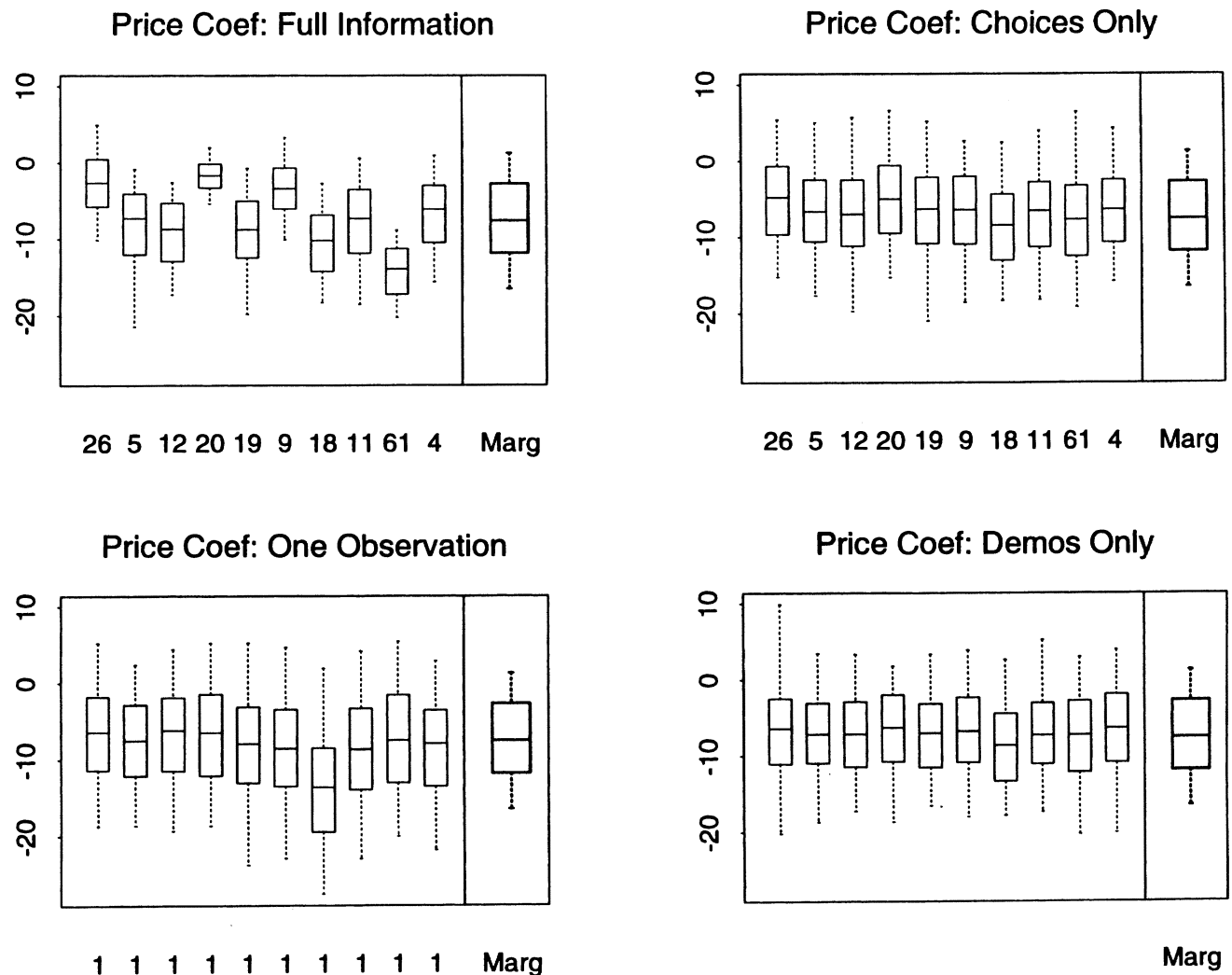
Price Coef: Full Information

Price Coef: Choices Only

Price Coef: One Observation

Price Coef: Demos Only

Figure 2 presents the marginal posteriors for the price coefficient for each of the five information sets. This figure is laid out in the same format as Figure 1. Here we see the tremendous value of the full information set in determining the level of price sensitivity. It is remarkable that even one observation (of both choice and causal variables) can yield some information about price sensitivity where as the Choices-Only and Demographics-Only information sets show little

information. In the Choices-Only set, demographic variables and the correlation between intercept and price sensitivities are the basis for information about price sensitivity.

Our preliminary analysis of the data suggests that there is substantial information in the individual household purchase histories which could be exploited for the purpose of customizing marketing activities. In addition, it appears that demographic information is of only

332

limited value. To put a specific value on these various information sets requires a concrete marketing problem. In §8, we develop a targeted couponing problem which we will use to value the information sets with a substantive metric.

7. Target Marketing

In this section, we introduce a target couponing problem as a metric by which to gauge the value of the various household information sets. As discussed above, the idea of customized couponing for grocery products is becoming more common. Point-of-purchase electronic couponing is now commonplace. Some chains (such as Dominick's Finer Foods, the second largest chain in the Chicago area) are implementing targeted feature advertising strategies via direct mail and abandoning the use of FSIs.

We introduce a target couponing problem which captures the essence of couponing behavior without complicating the basic choice model. We hope that this stylized couponing problem will give the reader a feel for the *potential* value of customized couponing strategies. We focus on the revenue side of couponing, without introducing the costs of issuing and redeeming coupons and implementing a customized coupon trigger strategy. Our view is that these costs are important, but changing for technological reasons, and that the most difficult problem from the point of view of the marketer is to measure the potential benefits from a targeted couponing strategy.

Our model is that the coupon acts as a temporary price cut equal to its face value. We take the perspective of a manufacturer determining optimal couponing strategy for a brand. The manufacturer focuses on the value of couponing to achieve brand switching, which is compatible with our brand choice model formulation. We ignore dynamic effects induced by consumer stockpiling in response to coupon availability. We assume that consumers receive the target coupons and retain them for possible use at the next purchase occasion. Current electronic couponing schemes marketed by Catalina Marketing Inc., issue coupons at the point of purchase for redemption at a future time. Other electronic couponing vendors issue coupons upon entrance to the store for redemption during that store visit. With current practices, there will be some coupon loss between the point of issue and the point of purchase. We do not attempt to assess this loss and include in our analysis.

In our choice model, the expected incremental sales generated from a coupon with face value F for the hth household is then:

$$\text{Incremental Sales} = \Pr(i|\beta_h, \Lambda, \text{price} - F, X)$$
$$- \Pr(i|\beta_h, \Lambda, \text{price}, X) \quad (17)$$

where $\Pr(i|\cdot)$ denotes the purchase probability for the ith alternative, β_h are parameters which describe household h's preferences and sensitivity to marketing variables, and X are nonprice covariates. Expected net revenue (π) is equal to:

$$\pi_F = \Pr(i|\beta_h, \Lambda, \text{price} - F, X)(M - F) \quad (18)$$

for a manufacturer margin of M and a coupon with face value F. Given β_h, our goal is to find the face value which will maximize π. For realism, we restrict attention to face value in multiples of 5 cents. We note that the optimal face value could easily be zero, which would mean that we do not trigger a coupon to this household, an option not available in the standard blanket couponing strategies.

As discussed above, any successful customization approach must deal directly with the problem of partial information and take parameter uncertainty into account in the decision problem. An extremely naive approach would be to insert the parameter estimates directly into the revenue maximization problem posed above. Because of nonlinearities in the profit function, this would result in an "over-confident" strategy which would make more extreme offers than warranted by the data information. By taking a full decision-theoretic approach, we avoid these difficulties altogether.

To incorporate parameter uncertainty, we consider the posterior distribution of expected household net revenue conditional on a given face value, F. We choose the face value F so as to maximize the value of expected net revenues averaged over the distribution of β_h as dictated by standard decision theory. Uncertainty in β_h comes into the problem through the posterior distribution of the choice probabilities.

Figure 3 presents the posterior distribution of choice probabilities for brand 1 (C-O-S water) conditional on average prices for three selected households. The posterior distributions in Figure 3 are based on the full information set, including the household purchase history. The posterior distributions show the heterogeneity in the three households in terms of expected probability for brand one as well as uncertainty regarding this probability. In addition, the highly nonnormal shapes of these posterior distributions demonstrate that large sample approximations would be inappropriate here.

To illustrate the optimal decision theoretic choice of coupon value, Figure 4 plots the posterior distribution of net revenues for various coupon face values for a specific household. We are plotting the posterior distribution of

$$R(\beta_h, \Lambda \mid F) = \Pr(i \mid \beta_h, \Lambda, \text{Price} - F, X)(M - F).$$

In our analysis, we assume that M (the manufacturers margin) is \$0.35. The household examined in Figure 4 is a household with only four choice observations. The substantial uncertainty in choice probabilities is reflected in the widely dispersed posterior distributions. The posteriors are displayed with somewhat nonstandard boxplots. The "dots" in each box plot are the value of revenue evaluated at "plug-in" estimates of the choice model parameters, $R(\beta_h = E[\beta_h], \Lambda = E[\Lambda] \mid F)$. The posterior means are the Bayes estimates of the parameters. Due to the nonlinearity of the net revenue function, R, the $E[R(\beta_h, \Lambda \mid F)] \neq R[E[\beta_h], E[\Lambda])$. Insertion of the plug-in estimates results in an overconfident view of the choice probabilities which produces an overestimate of the ex-

Figure 3 **Posterior distribution of C-O-S water-packed choice probabilities. Three selected households (with the number of purchase history observations indicated below each histogram). Note that these are the 1st, 3rd, and 10th households as ordered in our dataset.**
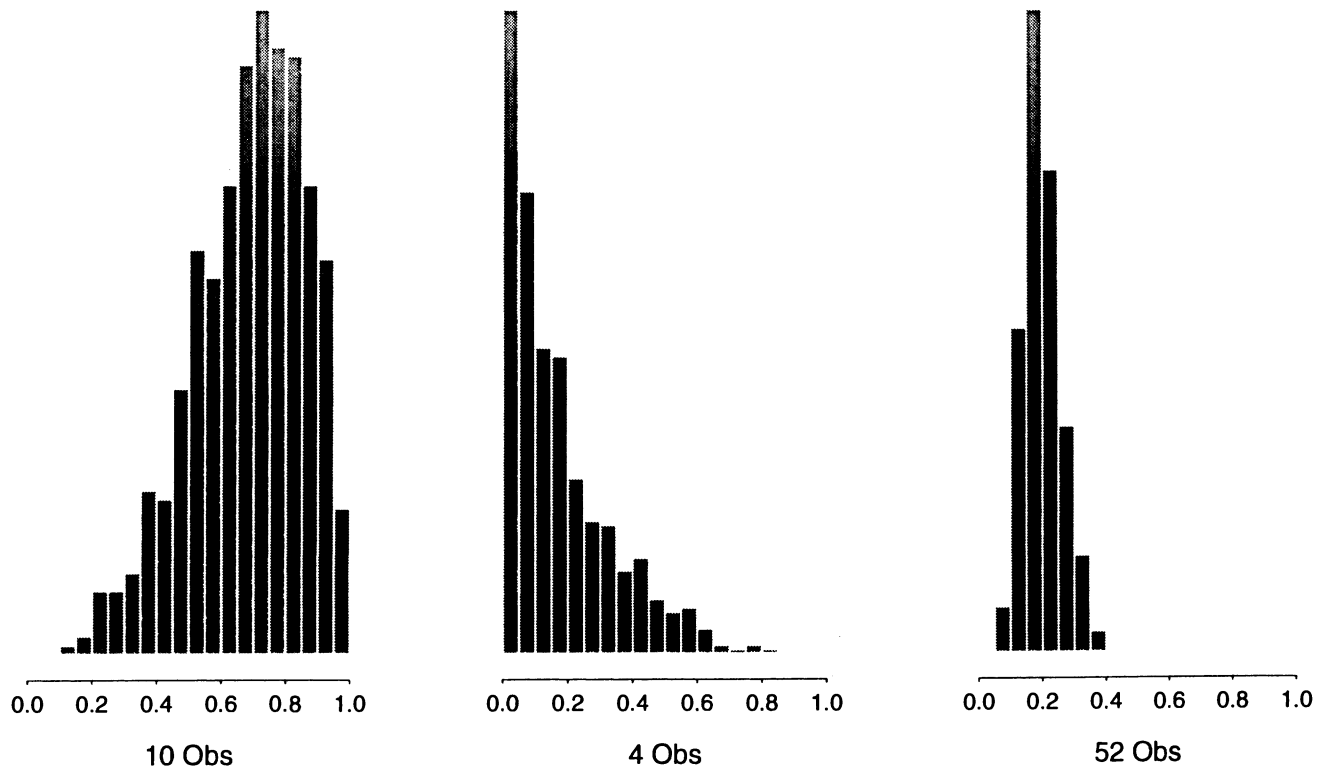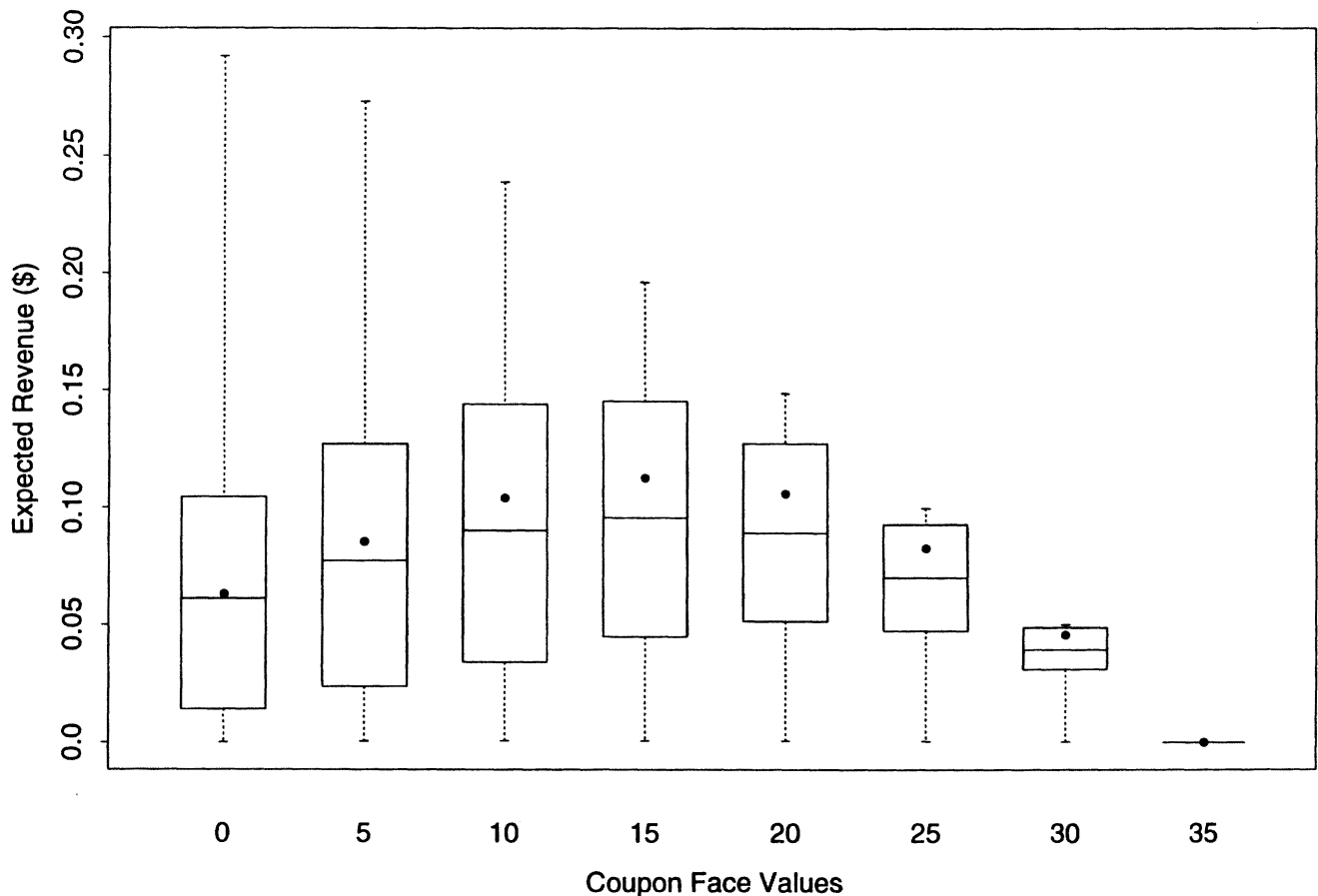

10 Obs

4 Obs

52 Obs

Figure 4 Posterior distribution of expected revenue for various coupon face values. One selected household. Expected revenue with parameters set equal to the posterior means ("plug-in") is shown by solid dots.

pected net revenues from couponing this household (all of the "dots" are above the means). For this household, both the full decision-theoretic solution and the "plug-in" solution give the same optimal face value, although this is not true for all households.

At the household level, it is important to use an approach which deals appropriately with the high level of uncertainty we can expect to encounter with only a limited number of observations. To evaluate the benefits from various different blanket and customized strategies, we must calculate the posterior distribution of aggregate profits and gauge whether there are sufficient gains and certainty to invest in the customization approach. This assessment is the goal of the next section.

8. The Value of a Household's Purchase History

The incremental value of household-specific information over base set of information can be calculated by comparing expected revenues realized from targeted vs. blanket couponing activities. A targeted coupon drop uses various household-specific information to infer about household preference parameters, and these inferences are used to compute an optimal customized face value of the coupon. An optimal blanket coupon drop uses information on the distribution of preferences in the population to design a single face value coupon.

As outlined in §4 above and illustrated in Figures 1 and 2, we construct marginal distributions of β_h

corresponding to each of the different information sets and then solve the optimal couponing problem for each information set. Figure 5 presents histograms of optimal C-O-S water brand face values for each of the information sets, based on a margin of 35 cents.

The dispersion of optimal face values is strikingly evident in Figure 5. For the full information set, we issue coupons as high as 30 cents in face value and issue no coupons at all to some $1/6$ of the sample. However, the "Demographics Only" dataset has almost no variability in optimal coupon face values. It is remarkable that the "One Observation" information set contains enough in-

formation to induce a fairly wide variation in coupon values. The "Choices-Only" information set induces a very similar, but not identical, pattern of wide variation in optimal face values.

The distributions of optimal face values are suggestive of a high value to customization, but a more definitive answer can be obtained from the aggregate distribution of net revenues. The aggregate net revenues are computed by summing up the profits from each household. Thus aggregate revenues are a function of all of the household-specific parameters, and to compute the posterior distribution of aggregate net revenue requires

Figure 5   Optimal Coupon Face Values. Various information sets.
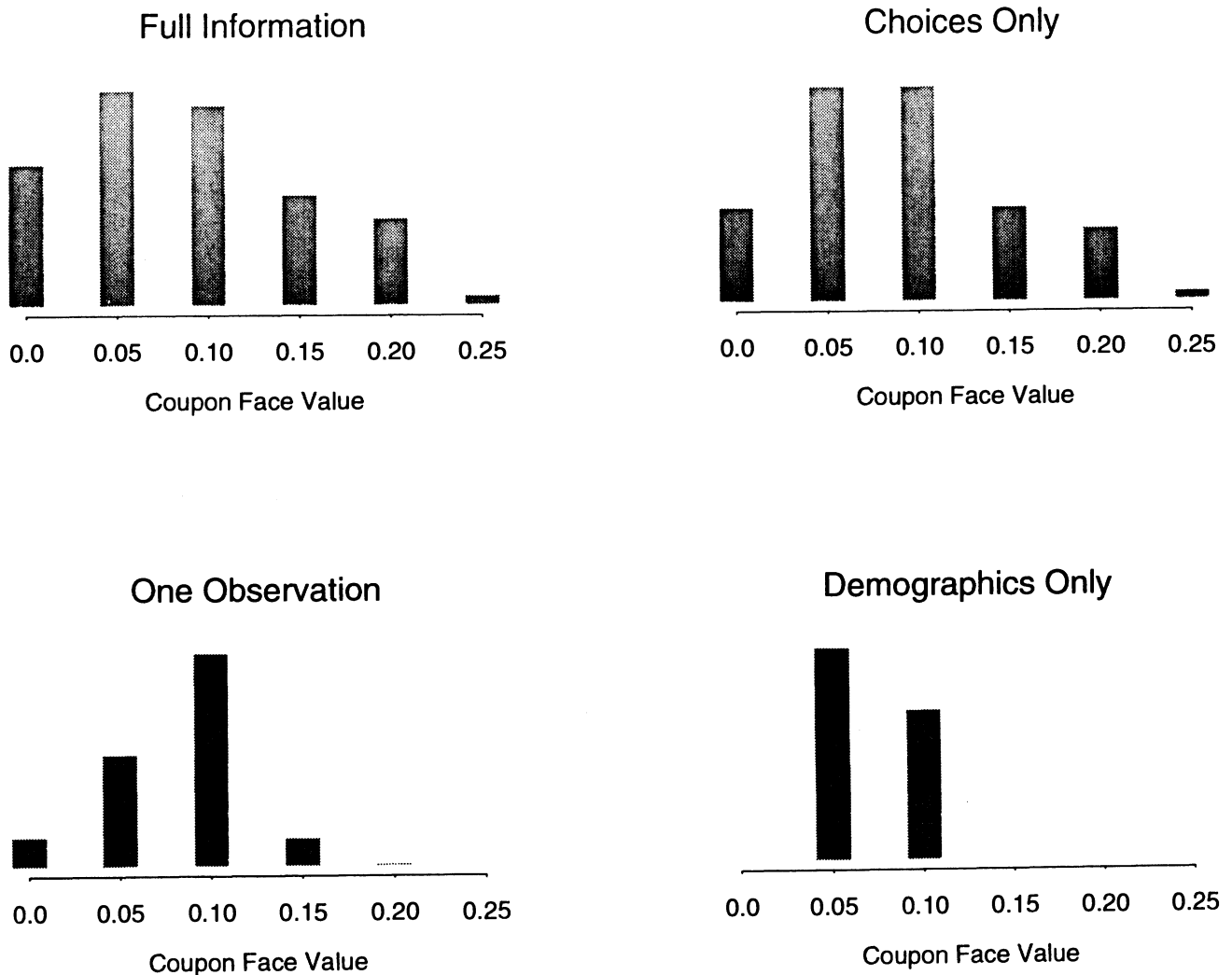
the joint distribution of the $\{\beta_h, h = 1, \cdots H\}$, which is a natural by-product of our probit sampling procedure. It should be noted that the β_h are not independent of each other since the hierarchical model pools information across households. Average net revenue (Π) is defined as follows:

$$\Pi(\beta_1, \ldots, \beta_h, \ldots, \beta_H, \Lambda)$$

$$= \Sigma_h \pi(\beta_h, \Lambda | F = F_h^*) / H. \quad (19)$$

We compute the posterior distribution of aggregate net revenue per household based on the optimal value values for each information set and evaluated with respect to the full information distribution of $\{\beta_h, h = 1, \ldots, H\}$ and Δ. These results are presented in Table 5. The last two lines of the table are added to consider the Aggregate Net Revenues generated by no coupon activity at all and the revenues generated by an optimal "blanket" strategy in which all households receive the same face value coupons. The last column labelled "Increment Over the Blanket" is the ratio of the increment in net revenue over blanket to the increment over no coupon, e.g., for Full—2.55 = (0.157 − 0.146)/(0.146 − 0.139).

We see that the gains from customization are potentially larger than the gains to blanket couponing in the first place (the full information gain is 2.5 times the blanket gain). While the Choices-Only information set is not as valuable as the full information, there is a substantial gain of as much as double the gains to blanket couponing. Even the "One Observation" information (on both choice and causal variables) set registers a substantial 50% gain over blanket couponing.

9. Concluding Remarks

The widespread use of optical scanning and other automatic data collection methods opens the possibility of collecting purchase history data for a substantial portion of the customer base. Many retailers have established loyalty or frequent shopper programs which depend on the collection of household-specific purchase history information. Electronic couponing schemes, such as those pioneered by Catalina Marketing, in which coupons are triggered directly by observed purchase behavior, are become increasingly common. Our conjecture is that the exponential decline the cost of information technology will enable the collection of much

Table 5   Relative Value of the Information Sets

| Information Set	Net Revenue	Gain Relative to Blanket
Full	0.1570	2.55
Choices-Only	0.1529	1.93
One Obs	0.1500	1.56
Demos-Only	0.1467	1.12
Blanket	0.1459	1.0
No Coupon	0.1380	

more comprehensive and detailed household information in the near future.

The goal of this paper is to investigate the value of various forms of household purchase data in terms of a specific target couponing problem. We find that the potential value of household purchase data is very substantial. Even rather short purchase histories (in our sample the average number of purchases 13.23) can produce a net gain in revenue from target couponing which is 2.5 times the gain from blanket couponing. The most popular current electronic couponing trigger strategy uses only one observation to customize the delivery of coupons. It was a surprise to use that even the information contained in observing one purchase occasion boasts net couponing revenue by 50% more than that which would be gained by the blanket strategy. It should be noted that our customized strategies allow for any coupon face value (in multiples of 5 cents), while the current Catalina strategy has only two face values (0 and some fixed amount). Thus, the current Catalina strategy may result in substantially less gain over a blanket method.

The popular frequent shopper or loyalty programs typically collect information only on what the consumer buys and not on the causal environment at the time of purchase. This sort of loyalty data has a great (and largely unexploited) potential value in targeted couponing. Our net revenue results show that purchase history data on Choices Only results in nearly as large a gain as the blanket coupon.

Finally, we want to emphasize that our focus is on the potential benefits from household purchase data, and we do not attempt to quantify many of the costs. It may well be that the cost of implementing target couponing (data storage, computing, clearing coupons and the

Figure 6   Comparison of finite and continuous mixture household parameter estimates (posterior means). Leftmost plots show the marginal distribution across households of the finite mixture estimates. Rightmost plots show the marginal distribution of household estimates for the continuous mixture model. Middle column of plots show a scattergram of finite mixture vs. continuous mixture estimates (solid line is the 45 degree line thru the origin).

C-O-S Oil Int

Price Coef
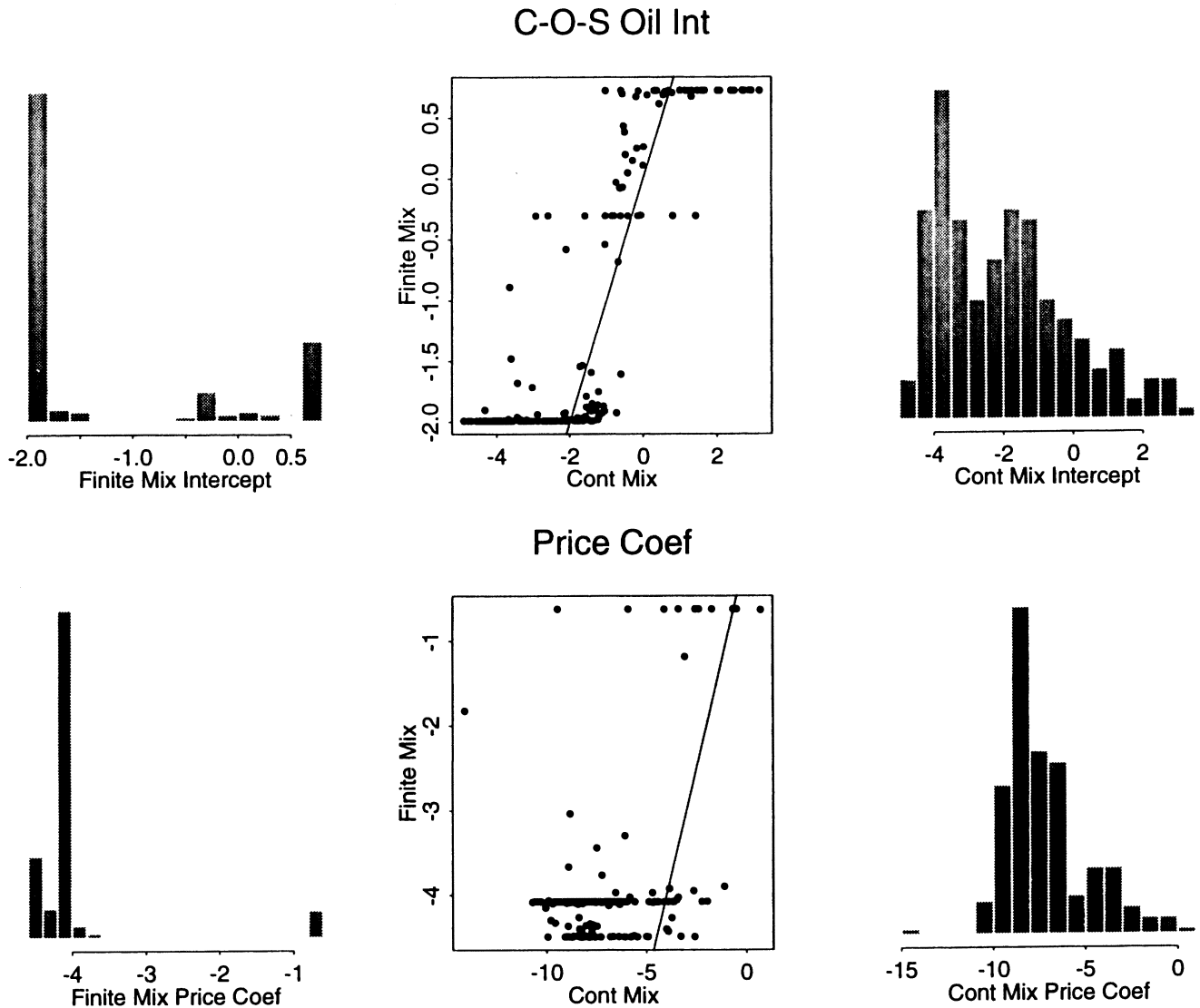
like) exceeds the benefits enumerated here. The costing of electronic couponing is more of an industry problem which is apt to become quickly out-dated as new technologies become available. In addition, our model of response to couponing may be simplistic, and we have not included effects such as gaming on the part of consumers which may reduce the potential benefits.

Appendix A. Priors and the Gibbs Sampler

Priors

There are three priors which are used in our hierarchical model: (a) the prior on Λ (the error variances in the random utility model); (b) the prior on V_β the covariance matrix of β_h given z_h; and (c) the prior on Δ, the matrix of regression coefficients in the model of heterogeneity, $\beta_h = \Delta z_h + v_h$.

(a) Prior on Λ

338

$$\Lambda = \text{diag}(\sigma_1^2, \ldots, \sigma_m^2),$$

$$\sigma_i \sim \text{independent Inverted Gamma}(\nu, \sqrt{v_i})$$

$$i = 2, \ldots, m \ (\sigma_1 = 1.0).$$

In our analysis, $\nu = 3$ and $v_i = 1.0$.

(b) Prior on V_β

$$V_\beta^{-1} \sim \text{Wishart}(\nu_{bo}, V_{bo}).$$

In our computations, $\nu_{bo} = k + 4$ (11) and $V_{bo} = \nu_{bo}I_k$. These prior settings keep the prior on V_β proper but quite diffuse. This means that there will be little shrinkage of the β_h toward the common subspace Δz_h.

(c) Prior on Δ

$$\delta = \text{vec}(\Delta) \sim N(\bar{d}, (V_\beta \otimes A_d^{-1})); \text{ this is the natural conjugate}$$

prior for the multivariate regression model.

In our computations, $\bar{d} = 0$, $A_d = 0.01I_d$.

Conditional Posteriors

The probit sampler we use is a Gibbs sampler which cycles through five sets of conditional posteriors.

I. $y_{h,t}|I_{h,t}, \beta_h, \Lambda, \Delta, V_\beta, h = 1, \ldots, H \quad t = 1, \ldots, T_h$.

Each of the households is conditionally (conditioned on X_h, z_h) independent, and each household observation is independent given β_h. This means that we can cycle down the huge vector of $\{y_{h,t}\}$ which is of length $\Sigma_h T_h$ by handing each $y_{h,t}$ separately. To reduce the notational burden, we drop the h and t subscripts. We now must draw from

$$y|I, \beta, X, \Lambda$$

where $y = X\beta + \epsilon \quad \epsilon \sim N(0, \Lambda)$, y is m dimensional.

This is a truncated m dimensional multivariate normal distribution

$$y \sim \text{truncated } N(\mu, \Lambda); \quad \mu = X\beta,$$

where the truncation is such that if $I = j$, then $y_j > y_k \quad k \neq j$.

As in McCulloch and Rossi (1992), we "Gibbs-thru" the vector y by drawing successively from m truncated univariate normals as follows:

start with $k = 1$,

if $k = I$,

draw y_k from $TN(\mu_k, \sigma_k, y_k > \max(y_{-k}))$

else

draw y_k from $TN(\mu_k, \sigma_k, y_k < \max(y_{-k}))$

increment k and return to top.

II. $\beta_h|y_{h,t}, \Lambda, \Delta, V_\beta$.

This is a standard Bayesian analysis of a linear regression with known residual covariance matrix and a normal prior. We first standardize the X's and y by premultiplying by Λ^{-1}

$$\beta_h \sim N(\bar{b}, (X_h'X_h + V_\beta^{-1})^{-1})$$

where X_h is a $mT_h \times k$ matrix of the stacked X_{ht} and

$$\bar{b} = (X_h'X_h + V_\beta^{-1})^{-1}[X_h'X_h\hat{\beta}_h + V_\beta^{-1}\bar{\beta}_h],$$

$$\bar{\beta}_h = \Delta z_h,$$

$$\hat{\beta}_h = (X_h'X_h)^{-1}X_h'y_h.$$

III. $\Lambda|y, \{\beta_h\}$.

$$\sigma_i \sim \text{Inverted Gamma}(\nu + n, \text{sqrt}((\nu v_i + ns_i^2)/(\nu + n))),$$

ns_i^2 is residual sum of squares

formed from the stacked vector of $\{e_h = y_h - X_h\beta_h\}$.

IV. $\Delta|\{\beta_h\}, V_\beta$.

$$\delta = \text{vec}(\Delta) \sim N(\bar{d}, (V_\beta \otimes (Z'Z + A_d)^{-1}))$$

where $\bar{d} = \text{vec}(\tilde{D})$, $\tilde{D} = (Z'Z + A_d)^{-1}(Z'Z\hat{D} + A_d\bar{D})^{-1}$,

$$\hat{D} = (Z'Z)^{-1}Z'B,$$

B is the $H \times k$ matrix with each β_h' as a row,

Z is the $H \times d$ matrix with each z_h' as a row,

$\bar{D} = \text{stack}(\bar{d})$, the $d \times k$ matrix formed

column by column from the elements of $\bar{d}$.

V. $V_\beta|\{\beta_h\}, \Delta$.

$$V_\beta^{-1} \sim W(\nu_{bo} + H, V_{bo} + S),$$

$$S = \Sigma_h(\beta_h - \bar{\beta}_h)(\beta_h - \bar{\beta}_h)', \quad \bar{\beta}_h = \Delta z_h.$$

Appendix B. Comparison with a Finite Mixture Approach

In order to facilitate a direct comparison between the continuous and discrete approaches to modeling heterogeneity, we implement a finite mixture approach for the independence probit model. As is standard in this literature, we use a BIC or Schwarz criterion method to identify the number of mass points.

Mass Pts	# Parms	logLike	BIC
1	13	−5440.8	−5496.5
2	21	−4657.5	−4747.5
3	29	−4436.2	−4560.5
4	37	−4418.5	−4577.1

We choose to report results for a three-mass-point model. We do not introduce demographic variables into the model, since the demographic variables have already been revealed to yield very little information. In addition, the most directly comparable way to add the demographic variables into the finite mixture model would be to interact them with all of the causal variables and intercepts which would have created a proliferation of parameters.

We compare the finite and continuous mixture models in terms of estimation of household level parameters. For the finite mixture model, we use the approach of Kamakura and Russell (1989) to compute an approximation to the posterior mean of β_h for each of the first 200 households in the data. We then evaluate the log-likelihood of the independence probit at the finite mixture posterior means of households parameters and compare this to the log-likelihood value for the posterior means of household parameters computed from our continuous model of heterogeneity. There is almost a 100% improvement in log-likelihood as a measure of goodness of fit from the continuous mixture approach (-1170 for our continuous mixture vs. -2036 for the finite mixture approach).

Figure 6 illustrates some of the reasons for the dramatic improvement in fit with the continuous mixture. In the top panel, we plot the posterior means for the finite and continuous mixture for the COS Oil Intercept. The leftmost graph shows the marginal distribution of finite mixture means. The posterior means for the finite mixture model must lie in the simplex defined by the three mass points. No one household can be any more or less "oil loving" that the extremes defined by two of the mass points. The support of the continuous mixture posterior means is not truncated so that we see households that are extremely oil or water loving. The scatter plot in the middle shows this phenomenon very clearly. The bottom three graphs show the same situation for the price coefficient. For the price coefficient, the finite mixture model makes few distinctions among the price sensitive households, while establishing a segment of households that is virtually completely price insensitive.

In this application to estimating household level parameters, the continuous mixture model provides a better fit to the data and more reasonable inferences.[1]

References

Allenby, Greg M. and James L. Ginter (1995), "The Effects of In-Store Displays and Feature Advertising on Consideration Sets," *International Journal of Research in Marketing*, 12, 67–80.

Borsch-Supan, A. and V. Hajivassiliou (1992), "Health, Children, and Elderly Living Arrangements: A Multiperiod-Multinomial Probit Model with Unobserved Heterogeneity and Autocorrelated Errors," in David Wise (Ed.), *Topics in the Economics of Aging*, Cambridge: NBER.

Bucklin, R. and S. Gupta (1992), "Brand Choice, Purchase Incidence, and Segmentation: An Integrated Approach," *Journal of Marketing Research*, 29, 201–215.

Chamberlain, Gary (1986), "Panel Data", in Zvi Griliches and Michael Intriligator (Eds.), *Handbook of Econometrics*, chapter 22.

Gelman, A. and D. Rubin (1992), "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7, 457–511.

Gelfand, Alan E. and Adrian F. M. Smith (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelfand, A. et al. (1990), "Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling," *JASA*, 85, 972–985.

Gupta, S. and P. K. Chintagunta (1994), "On Using Demographic Variables to Determine Segment Membership in Logit Mixture Models," *Journal of Marketing Research*, 31, 128–136.

Hajivassiliou, V. (1993), "Simulation Methods for Limited Dependent Variable Models," forthcoming, *Handbook of Statistics*.

Hausman, J. and D. Wise (1978), "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," *Econometrica*, 46, 403–426.

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

Keane, M. (1994), "Simulation Estimation for Panel Data Models with Limited Dependent Variables," in C. R. Rao, G. S. Maddala and H. D. Vinod (Eds.), *The Handbook of Statistics 11: Econometrics*, Amsterdam: Elsevier Science Publishers, 545–572.

McCulloch, R. and P. Rossi (1994), "An Exact Likelihood Approach to Analysis of the MNP Model," *Journal of Econometrics*, 64, 207–240.

McFadden, Daniel (1974), "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembda (Ed.), *Frontiers in Econometrics*, New York: Academic Press.

Tanner, T. and W. Wong (1987), "The Calculation of Posterior Distributions by Data Augmentation," *JASA*, 82, 528–49.

Tierney, L. (1991), "Markov Chains for Exploring Posterior Distributions," Technical Report No. 560, School of Statistics, University of Minnesota, Minneapolis, MN.