Direct Elicitation of Indirect Preferences

Daniel McFadden
University of California - Berkeley

Follow this and additional works at: http://repository.cmu.edu/sem_conf

Part of the Economics Commons

http://repository.cmu.edu/sem_conf/2015/full_schedule/144

This Event is brought to you for free and open access by the Conferences and Events at Research Showcase @ CMU. It has been accepted for inclusion in Society for Economic Measurement Annual Conference by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.
FOUNDATIONS OF STATED PREFERENCE ELICITATION
CONSUMER CHOICE BEHAVIOR, MEASUREMENT OF CONSUMER WELFARE, AND CHOICE-BASED CONJOINT ANALYSIS

Moshe Ben-Akiva, Daniel McFadden, and Kenneth Train

July 4, 2015

Abstract: This paper re-examines the discrete choice methods and stated preference elicitation procedures that are commonly used in choice-based conjoint analysis. The aim is to clarify their domains of applicability and provide reliable techniques for data collection and analysis.

CONTENTS

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>2</td>
</tr>
<tr>
<td>1 Choice Behavior and Consumer Welfare</td>
<td>2</td>
</tr>
<tr>
<td>1.1 Utility</td>
<td>5</td>
</tr>
<tr>
<td>1.2 Choice Probabilities</td>
<td>8</td>
</tr>
<tr>
<td>1.3 Consumer Welfare</td>
<td>10</td>
</tr>
<tr>
<td>1.4 Estimation</td>
<td>12</td>
</tr>
<tr>
<td>2 Some History of Stated Preference Elicitation</td>
<td>13</td>
</tr>
<tr>
<td>3 Choice-Based Conjoint Analysis (CBC)</td>
<td>15</td>
</tr>
<tr>
<td>3.1 Sampling and Recruitment</td>
<td>18</td>
</tr>
<tr>
<td>3.2 Design of Conjoint Experiments</td>
<td>19</td>
</tr>
<tr>
<td>3.3 Subject Training</td>
<td>21</td>
</tr>
<tr>
<td>3.4 Incentive Alignment</td>
<td>22</td>
</tr>
<tr>
<td>3.5 Reconciliation and Validation</td>
<td>23</td>
</tr>
<tr>
<td>3.6 A Conjoint Study Example</td>
<td>23</td>
</tr>
<tr>
<td>3.7 Stated Perceptions, Expectations, and Well-Being</td>
<td>25</td>
</tr>
<tr>
<td>4 Maximum Simulated Likelihood (MSL) Analysis of CBC Data</td>
<td>26</td>
</tr>
<tr>
<td>5 Hierarchical Bayes (HB) Estimation</td>
<td>29</td>
</tr>
<tr>
<td>5.1 Bayesian Estimation</td>
<td>30</td>
</tr>
<tr>
<td>5.2 Selecting Priors and Computing Posteriors</td>
<td>36</td>
</tr>
<tr>
<td>5.3 Hierarchical Bayes Methods</td>
<td>43</td>
</tr>
<tr>
<td>5.4 Policy Simulations</td>
<td>46</td>
</tr>
<tr>
<td>5.5 A Monte Carlo Example</td>
<td>47</td>
</tr>
<tr>
<td>6 An Empirical CBC Study using MSL and HB Methods</td>
<td>51</td>
</tr>
<tr>
<td>7 Caveats and Conclusions</td>
<td>55</td>
</tr>
<tr>
<td>References</td>
<td>56</td>
</tr>
<tr>
<td>Appendix: R and RSTAN code</td>
<td>63</td>
</tr>
</tbody>
</table>
Foundations of Stated Preference Elicitation

Moshe Ben-Akiva, Daniel McFadden, and Kenneth Train

Preface

Information on consumer preferences and choice behavior can be used to forecast market demand for new or modified products, estimate the effects on consumer welfare of market changes, develop and test models of consumer behavior, and reveal determinants and correlates of tastes. Direct elicitation of stated preferences, perceptions, expectations, choices, and well-being, supplementing or substituting for information on revealed choices in markets, can be a valuable source of information on consumers. The purpose of this paper is to provide the reader with the theoretical, statistical, and behavioral underpinnings of methods for direct preference elicitation. One focus will be on the economic theory and statistical analysis of choice behavior, revealed or stated, and an economic framework for forecasting market demand and measurement of consumer welfare. Another will be on the collection of preference and related data from consumer responses in hypothetical choice experiments, particularly conjoint analysis methods that have proven useful in market research. We concentrate on predictable choice behavior in markets for consumer goods and services, and we emphasize methods that can provide reliable, accurate forecasts. This endeavor is informed by and benefits from investigations of perceptions and decision-making behavior in cognitive science and behavioral economics, but this paper will not focus on stated preference experiments designed to reveal cognitive and behavioral characteristics of individuals.

There are a number of good introductions to discrete choice analysis (Ben-Akiva and Lerman, 1985; Train, 1986, 1993; McFadden, 1999, 2001, 2014; Brownstone, 2001; Boyce and Williams, 2015), and to stated preference and conjoint analysis methods applied to market research (Louvier, 1988; Louvier-Hensher-Swait, 2000; Hauser and Rao, 2002; Rossi-Allenby-McCulloch, 2005; Raghavarao-Wiley-Chitturi, 2010; Rao, 2014). This paper will complement these introductions by filling in and updating the technical and behavioral background of these subjects.

1. Choice Behavior and Consumer Welfare

The neoclassical economic theory of consumer choice behavior is based on the idea that consumers make choices to maximize preferences, subject to budget constraints determined by income and prices. Preferences are represented by utility functions defined over bundles of continuous and discrete market goods. Utility maximization can be decomposed into two stages: In the first stage, given each possible discrete alternative, a conditional maximum utility is obtained by optimization over the remaining market goods subject to the budget constraint. In the second stage, a choice is made from the menu of available discrete alternatives that maximizes these conditional
maximum utilities. McFadden (1974), Deaton and Muelbauer (1980), Fosgerau et al. (2013), and McFadden (2014) summarize the properties of these optimization stages. Different consumers can have different preferences, leading to distributions of market good demands and discrete choice probabilities over the population. However, a single consumer’s choices from different menus, which may be offered at different times or in different stages of a stated choice experiment, will in the neoclassical theory be explained by maximization of a single preference order.

Consumer choice from repeated menus in laboratory and market experiments often deviates from the strict neoclassical theory of choice; see McFadden (1999). There are several possible reasons: First, preferences may encompass unmeasured attributes of alternatives, and may be conditioned on previous choices, experience, and social motivations that are not fully measured and accounted for. Second, consumers may be poor statisticians, with inconsistent perceptions and expectations. Third, consumers may not be relentless and exact preference maximizers, instead making choices that are influenced by whimsy, fatigue, inattention, bias toward immediate gratification, and a willingness to settle for “good enough”. Fourth, consumers may have trouble discriminating among very similar alternatives, particularly if they have ambiguous or shrouded attributes.

The early models of individual choice behavior that came out of psychophysics, Thurstone (1927), Luce (1959), and Marschak (1960), focused on stochastic elements in individual preferences originating from imperfect psychophysical discrimination. When McFadden (1974) adapted these models for econometric analysis of choice behavior in markets, both inter-consumer and intra-consumer sources of preference heterogeneity were recognized, but lumped together since they are indistinguishable in surveys that collect data on a single choice by each subject. However, when considering stated choice data from subjects confronted with a panel of menus, intra-consumer and inter-consumer preference heterogeneity can be identified, and the question will be whether it is theoretically and statistically useful to do so.

One modelling strategy is to simply ignore the panel data structure established by multiple menu choices of multiple consumers, and treat each observed choice as the result of an independent draw from a distribution of preferences. This loses information and statistical efficiency when there is little intra-consumer heterogeneity, but is nevertheless a consistent model that in many applications will have good predictive power for market behavior. Another strategy is to retain most of the neoclassical assumption of stable preferences within consumers, with minor allowance for the effects of imperfect psychophysical discrimination. This fully exploits the information contained in multiple choices, and is particularly valuable if a target of the analysis is recovery of information on individual preferences or on the dynamics of choice behavior. However, it can give a misleading picture of individual

---

2 There are testable conditions from the theory of revealed stochastic preference on choice probabilities in a population of preference maximizers; McFadden (2006) summarizes the literature. For example, the probability of a specified alternative being chosen cannot rise when a new alternative is added to the menu, and the sum of the probabilities of choices in an intransitive cycle of length K cannot exceed K-1. When these testable conditions are satisfied, one cannot tell whether choice probabilities come from intra-personal or inter-personal heterogeneity.
preferences if intra-consumer heterogeneity is in fact significant. A third strategy is to introduce a structural system with both intra-consumer and inter-consumer stochastic elements in preferences; see Rose et al. (2009), Hess and Train (2011), Yanez et al. (2011), and Daly-Hess-Stathopoulos (2012). In this paper, we adopt the second strategy, assuming that a consumer’s choices over multiple menus maximize the same preferences, aside from errors due to the limits of psychophysical discrimination. This setup also covers the first strategy through the expedient of treating multiple choices from one consumer as if they were single choices from multiple consumers. We will not pursue the third strategy in detail, but will indicate where and how it could be implemented.

Suppose consumer n in a sample indexed n = 1,…,N has socioeconomic characteristics $s_n$, including household demographics, location, and past market experience, and real disposable income $I_n$. It will often be important to decompose income into components, wage and salary income $W_n$, net asset income $A_n$, and net lump-sum transfer income $T_n$ (typically negative due to taxes); then $I_n = W_n + A_n + T_n$. The reason for this decomposition is that in a population with heterogeneous preferences, choice probabilities can vary with income for three reasons: First, a consumer with given preferences may shift consumption patterns as disposable income rises, a classical income effect. Second, changes in wage or asset income may be linked to the effective prices of discrete alternatives or attractiveness of features; e.g., through the opportunity cost of leisure or the cost of financing purchases of durables. Third, observed income may be ecologically correlated with tastes across the population due to common causes; e.g., consumers with high rates of impatience may have both lower incomes because of unwillingness to delay consumption, and have higher probabilities of discrete choices that have immediate payoffs, with the correlation coming from heterogeneity in impatience rather than a causal link from income to choice probability. In modelling utility, the first income effect should enter the individual utility model. The second effect should enter the determination of the consumer’s budget and effective prices of alternatives, but not the utility model, and should not involve transfer income. The third ecological effect should not be a factor in individual decision-making. Compensating variations to offset policy changes are income transfers that affect transfer income, but have no effect on wage or asset income. Further, ecological correlations of income and demand behavior should have no effect on individual compensating variation. Finally, when compensating variations are hypothetical, and not actually paid, as is the case in most applications of consumer welfare analysis, they can have no effect on actual disposable income $I_n$, and consequently on demand response to changes in actual disposable income. Then, to handle income effects correctly in consumer welfare calculations, it is necessary to avoid confounding income effects on individual decision-making with ecological correlations, confounding the effect of transfer income changes with changes in wage or asset income components, and confounding hypothetical and actual compensation.
1.1 Utility. Suppose that in each of \( m = 1, \ldots, M \) different menus, consumer \( n \) faces a discrete choice of whether to buy one of \( j = 1, \ldots, J_m \) alternative products, or buy nothing \( (j = 0) \). Suppose product \( j \) in menu \( m \) has a real price \( p_{jm} \) and a vector of observed “raw” attributes \( z_{jm} \). Suppose consumer \( n \) chooses among the alternatives \( j = 0, \ldots, J_m \) in menu \( m \) to maximize (indirect) money-metric or Willingness-To-Pay-space utility \(^3\)

\[
(1) \quad u_j = U(W_n, A_n, T_n, x_{jnm}, q_{jnm}, p_{jmn} | \alpha_n, \beta_n, \delta_n) \equiv v_{jmn} + \alpha_n \varepsilon_{jmn},
\]

where \( v_{jnm} = I_n - p_{jmn} + x_{jmn} \beta_n + q_{jmn} \delta_n \) is a dollar-denominated “systematic” component of utility, \( \alpha_n \varepsilon_{jmn} \) is a “stochastic” component that can vary across menus for the consumer, \( x_{jnm} = X(W_n, A_n, s_n, z_{jmn}) \) is a vector of known transformations (e.g., logs, powers) that can include nonlinear transformations and interactions of raw product attributes and interactions of raw product attributes and consumer characteristics, \( q_{jnm} = Q(z_{jmn}) \) is a vector of dummy variables that identify types or classes of products (e.g., indicators for “brand”), and \( p_{0mn} = 0, x_{0nm} = 0, q_{0nm} = 0 \) by definition. The \( x_{jnm} \) are weighted by parameters \( \beta_n \) that are interpreted as consumer \( n \)’s part-worth or Willingness-To-Pay (WTP) values for these attributes \(^4\), the \( q_{jnm} \) are weighted by factors \( \delta_n \) that are interpreted as idiosyncratic tastes for products or product classes, and \( \alpha_n \) is a positive parameter that scales the additive unobserved random disturbance \( \varepsilon_{jnm} \). The parameters \( \alpha_n, \beta_n, \) and \( \delta_n \) will in general be heterogeneous across consumers, and interpreted in empirical analysis as random effects. For brevity and later use in computational formulas, we introduce an intermediate parameter vector \( \nu_n \), write the parameters in (1) as known transformations \( \alpha_n = \alpha(\nu_n), \beta_n = \beta(\nu_n), \delta_n = \delta(\nu_n) \), and write \( V_{jnm}(\nu) \equiv I_n - p_{jmn} + x_{jmn} \beta(\nu) + q_{jnm} \delta(\nu) \) so that \( v_{jnm} = V_{jnm}(\nu_n) \).

Examples of \( X(W_n, A_n, s_n, z_{jmn}) \) transformations, from a study of automobile choice, are horsepower divided by vehicle weight and the number of seats in the vehicle less the number of persons in the consumer’s household. The \( \beta_n \) is a “representative taste” parameter vector for consumer \( n \) that is stable across menus. For example, if the products are automobiles and a component of \( x_{jnm} \) is fuel efficiency, then the corresponding component of \( \beta_n \) is interpreted as WTP for a unit increase in fuel efficiency; this measure reflects both tastes and the consumer’s

\(^3\) The function (1) has the Random Utility Model (RUM) form introduced by McFadden (1974) that is a common starting point for discrete choice analysis. His original distinction between systematic and stochastic components remains useful for analysis, but in the current specification, both components can include stochastic elements. The term “money-metric” utility is due to Samuelson (1950), who observed that indirect utility evaluated at given prices could with a monotone transformation be expressed in units of income. Hurwicz and Uzawa (1971) showed the connection of this form of utility to measurement of consumer surplus; see also McFadden (2014). The term “WTP-space” utility and recognition of its usefulness in stabilizing empirical measures of consumer welfare in discrete choice problems is due to Train and Weeks (2005); see also Daly-Hess-Train (2012), and Carson-Czajkowski (2013). Train and Weeks use the terminology “preference-space” utility when (1) is alternately scaled so that the coefficient on the stochastic component is one.

\(^4\) WTP for an incremental change in the vector of raw attributes \( z_{jmn} \) is given by \([X(s_n, z_{jmn} + \Delta z_{jmn}) - X(s_n, z_{jmn}) \beta_n] \). This change may be confined to a single component of \( z \), or span several components, and any change may be confined to a single component of \( X \) for a single alternative, or may span multiple components and alternatives.
subjective expectations on the future price of fuel. The $x_{ijmn}$ may also depend on wage and asset components of income. For example, in a study of travel mode choice, a component of $x_{ijmn}$ might be hours of travel time multiplied by the decision-maker’s real wage. Examples from automobile choice are an interaction of income component $W_n$ with tastes for time-saving attributes, and interaction of net asset income with an attribute related to collateral requirements for financing. In these examples, $x_{ijmn}$ does not depend on net transfer income, this preserves the WTP interpretation of the $\beta_n$. The $q_{ijmn}$ can include dummy variables that are turned on for alternatives that share specific characteristics; e.g., dummy variables for “brand name” or for commodity classes (e.g., “buses” or “public transit” in a study of mode choice), with the associated components of $\delta_n$ capturing the common impact of these factors on the utilities of various choices.

An issue that is important later for predicting demand for new products is whether $x_{ijmn}$ and $q_{ijmn}$ depend only on “generic” attributes of products, or also on “name-specific” attributes. Here, “generic” means that consumers respond to the attribute in a new product in the same way as in an existing products; e.g., “schedule delay time” (the time interval between when the consumer would like to leave and when the mode is scheduled to leave) in a study of transportation mode choice has the same definition for both existing and new modes. Conversely, “name-specific” means an attribute that interacts with a “named” class dummy, so that the associated WTP can depend on the “name” of the product. For example, WTPs may differ for “airline schedule delay time” and “high-speed rail schedule delay time”. Note that class indicators $q_{ijmn}$ may be either generic (e.g., “public transit”) or name-specific (e.g., “ABC Bus Company”). While name-specific attributes may provide a parsimonious “explanation” of tastes, they are short-hand for deeper generic attributes, such as the comfort of the waiting environment in the mode choice example. These short-hand variables may be convenient, or expedient when it is clear how to extend name-specific attributes to new products; e.g., when brand-specific effects are not easily characterized in terms of perceived reputation, or when new products are introduced as new lines in an existing brand. However, the analyst should recognize that characterizing “brand” in terms of generic attributes such as reputation for reliability or durability extends the ability of analysis to forecast demand for innovative products.

The disturbances $\varepsilon_{ijmn}$ are interpreted as perturbations in utility that will vary from one consumer, menu, and alternative to the next, capturing the effects of variations in WTP for specific products and product attributes. These perturbations may come from whimsy, from the psychometric difficulty consumers have in distinguishing between products and attribute levels, from lack of consumer attention and acuity, and from mental mechanisms consumers

---

5 When individual disposable income, including transfer income, influences choice because of a classical income effect, there is a theoretical question as to whether hypothetical compensating variations to offset policy changes should be interpreted as an integral change in disposable income that influences choice probabilities, or interpreted as adjustments after-the-fact that do not influence demands. In the former case, the simple WTP interpretation of the $\beta_n$ is lost, but it is still possible to calculate exact WTP for an incremental change in raw attributes using a one-dimensional non-linear search; see Dagsvik and Karlstrom (2005), McFadden (2014).
use to “break ties” when the utility of alternatives seem indistinguishable. It is important to recognize that while the perturbations $\varepsilon_{jmn}$ may be needed to reconcile observed choices with utility maximization, they are not an “explanation” of choice behavior. If their presence looms large in rationalizing choice, this says that a lot of the determinants of choice are unobserved and uncontrolled, and this will limit the accuracy of predictions. Ideally the $\delta_n\varepsilon_{jmn}$ terms should have a modest effect on utility differences relative to the “systematic” component $\nu_{jmn}$ that contains the available information on the consumer’s choice environment. When this is true, it is possible to view the $\varepsilon_{jmn}$ as elements that simply smooth responses and facilitate computation, so that predictions of choice will be relatively insensitive to the specification of their distribution. On the other hand, when choice is poorly predicted by the systematic component, specification of the distribution of $\varepsilon_{jmn}$ will have a significant impact on predicted choice probabilities, and it will become important to avoid misspecifications that lead to unrealistic choice probabilities. When there is a sufficiently rich array of dummy variables in $q_{jmn}$ whose coefficients $\delta_n$ capture idiosyncratic tastes for various product classes, the disturbances $\varepsilon_{jmn}$ are not needed to “explain” correlations between utilities of different alternatives, and there is little harm in assuming that they are independent.

The $\varepsilon_{jmn}$ are often assumed to be independently identically distributed with Extreme Value Type 1 (EV1) distributions. The additive linear specification of utility in (1) combined with the EV1 assumption on the $\varepsilon_{jmn}$ seems quite restrictive, but in fact (1) with sufficiently flexible specification of the $x_{jmn}$ and $q_{jmn}$ and inter-consumer heterogeneity in the parameters $\beta_r$ and $\delta_n$ can approximate any discrete choice behavior that is consistent with regular preference maximization. A substantive assumption that $X$ is globally independent of $T - p_j$ is common in applications and is often empirically reasonable, particularly when wage and asset income can enter $X$ separately. The

6 The CDF of an EV1 random variable $\varepsilon$ is $\exp(-\varepsilon^\alpha)$. From Johnson and Kotz (1970), Chap. 21, a linear transformation $\nu = \mu + \alpha \varepsilon$ has CDF $\exp(-\exp(-\nu^\alpha))$, mean $\mu + \alpha \gamma_0$, where $\gamma_0 = 0.57721$ is Euler’s constant, mode $\mu$, median $\mu - \alpha \log \log 2$, and variance $\alpha^2 \pi^2/6$. If $\varepsilon_j$ are i.i.d. EV1 for $j = 0, ..., J$, the CDF of $\nu = \max_{j=1,...,J}(\mu_j + \alpha \varepsilon_j)$ is $\prod_{j=1}^J \exp(-\exp(-\nu^\alpha)) = \exp(-\exp(-\nu^\alpha))$, where $\rho = \alpha \cdot \log(\sum_{j=1}^J \exp(\varepsilon_j^\alpha))$. Then $\nu$ has mean $\mu + \alpha \gamma_0$ and variance $\alpha^2 \pi^2/6$. The probability of $\mu_0 + \alpha \gamma_0 \geq \max_{j=1,...,J}(\mu_j + \alpha \varepsilon_j)$ is $\int_{\varepsilon_0=-\infty}^{\infty} \exp(-\exp(-\nu^\alpha)) \prod_{j=1}^J \exp(-e^{-\exp(\varepsilon_j^\alpha)}) d\varepsilon_0$. Collecting terms, this integral becomes $\int_{\varepsilon_0=-\infty}^{\infty} \exp(-e^{-\exp(\varepsilon_j^\alpha)} \cdot \prod_{j=1}^J [1 + \sum_{i=1}^J \exp((\mu_i - \mu_0)/\alpha)] \, d\varepsilon_0 = 1/[1 + \sum_{i=1}^J \exp((\mu_i - \mu_0)/\alpha)]$.

7 This result follows from McFadden and Train (2000), who consider any population of consumers with heterogeneous complete transitive preferences, indexed by elements $r$ in a compact space $R$, over objects $t$ in a compact space $T$. They first show (MT Lemma 1) that a mild preference continuity condition across outcomes and consumers implies that these preferences can be represented by a utility function $U$ that is continuous in $(t,r)$. Second, they show (MT Theorem 1) that when these consumers face a discrete choice over $j = 0, ..., J$ alternatives whose features $t = (t_0, ..., t_J)$ are finite-dimensional, and the probability of ties between the utilities $U(t_i,r)$ is zero, then for any small positive scalar $\eta$, there exists a vector of transformations $x_i = X(t_i)$ and a commensurate vector of parameters $\beta(r)$ such that utility function $U^*(t_i,r) = X(t_i)\beta(r)$ ranks alternatives differently than $U(t_i,r)$ with probability less than $\eta/2$. Then the utility function $X(t_i)\beta(r) + \alpha \varepsilon_i$, where the $\varepsilon_i$ are i.i.d. EV1 distributed, will for sufficiently small $\alpha$ rank alternatives differently than $U(t_i,r)$ with probability less than $\eta$. In our application to product choice, let $t_j = (W,A,T,p_s,z_0)$. Given a family of utility functions $U^*(W,A,T,p_s,z_0,r)$ that are strictly increasing in $W,A,T$ as a result of local non-satiation, note that $u = U^*(W,A,T,0,s_0,z_0,r)$ has a continuous inverse $T = M(u,W,A,0,s_0,z_0,r)$ and make a continuous increasing transformation $U(W,A,T,p_s,z_0,r) = M(U^*(W,A,T,p_s,z_0,r);W,A,0,s_0,z_0,r)$. Then $U$ has the money metric property that $U(W,A,T,0,s_0,z_0,r) = T$. Then this utility can be approximated by (1) with attributes $x_i = X(T - p_j,W,A,s_0,z_0)$ that are to first order independent of $T - p_j$. This result allows $\alpha$ be a fixed parameter that is homogeneous across the population.
arguments above absorb class dummies \( q_{mn} \) into \( X \), but they are separated out in (1) for ease of interpretation as carriers of inter-consumer heterogeneous class effects.

1.2 Choice Probabilities. To reduce notation, assume for the rest of this paper that \( M_n = M \) and \( J_{mn} = J \) are the same for all menus and consumers; the reader can reattach these subscripts as needed. Define \( x_{mn} = (x_{1mn}, \ldots, x_{Jmn}) \), \( q_{mn} = (q_{1mn}, \ldots, q_{Jmn}) \), and \( p_{mn} = (p_{1mn}, \ldots, p_{Jmn}) \). The next step in discrete choice analysis is to go from the utility (1) to a choice probability model consistent with its maximization, given observed \( x_{mn}, q_{mn}, p_{mn} \), and proximate or deep parameters. Suppose the disturbances \( \varepsilon_{jmn} \) are distributed i.i.d. Extreme Value Type 1 (EV1). From footnote 6, the probability that the consumer maximizes (1) with choice \( k \) is then of multinomial logit (MNL) form

\[
P_k(x_{mn}, q_{mn}, p_{mn} | \nu_n) = \frac{\exp\left(\frac{V_{kmn}(\nu_n)}{\alpha_n}\right)}{1 + \sum_{j=1}^{J} \exp\left(\frac{V_{jmn}(\nu_n)}{\alpha_n}\right)}
\]

for \( k = 0, 1, \ldots, J \).

In many applications, the taste parameters \((\alpha_n, \beta_n, \delta_n) = (\alpha, \beta, \delta)\) are assumed to be the same for all consumers. Then (2) reduces to the flat MNL model,

\[
P_k(x_{mn}, q_{mn}, p_{mn} | \alpha, \beta, \delta) = \frac{\exp\left(\frac{x_{kmn} \beta + q_{kmn} \delta - p_{kmn}}{\alpha}\right)}{1 + \sum_{j=1}^{J} \exp\left(\frac{x_{jmn} \beta + q_{jmn} \delta - p_{jmn}}{\alpha}\right)}
\]

for \( k = 0, 1, \ldots, J \),

A strong and potentially seriously limiting property of (3) is Independence of Irrelevant Alternatives (IIA): the odds that one alternative is chosen over a second are unchanged when a third alternative is added to the menu. When the new alternative shares many features with the second alternative but not the first (e.g., the first two alternatives are “car” and “red bus”, and the new alternative is “blue bus”), it may primarily split bus choices choice rather than drawing from the car choice. In this case, (3) will over-predict bus demand. This is not a serious issue if alternatives do not have features (e.g., “bus-i-ness”) in common that lead consumers to lump them together, or if observed choices mostly agree with the ranking of alternatives according to their systematic component of utility \( V_{jmn}(\nu) \). In general, IIA need not be a problem in the general model (2), where “class” dummies in \( q_{mn} \) and variations in \( \delta_n \) across consumers are usually able to track preference patterns (e.g., tastes for “buses”) closely.

When the IIA property is not reasonable, it is possible to alter the assumed distribution of the disturbances \( \varepsilon_{jmn} \), say by adopting a Multinomial Extreme Value distribution for these disturbances, but it is more general and usually more practical to adopt the model (2) that leaves the i.i.d. EV1 assumption on the \( \varepsilon_{jmn} \) in place, but specifies sufficiently expansive and flexible \( q_{mn} \) and \( \delta_n \) so that this model incorporates correlations between the utilities of different alternatives and comes close to fully “explaining” choice.

When a population of consumers is considered, inter-consumer heterogeneities in tastes are to be expected that will appear in (1) and (2) as heterogeneities in \((\alpha_n, \beta_n, \delta_n)\). Possible modelling strategies when confronted with choice data generated by (2) are (A) to impose homogeneous \((\alpha, \beta, \delta)\), obtaining the flat MNL model (3) in which any inter-
consumer heterogeneities that are present are “subsumed” in the \( \varepsilon_{jmn} \); (B) to allow \((\alpha_n, \beta_n, \delta_n)\) to be a separate “fixed effects” parameter vectors for each consumer; or (C) to specify \((\alpha_n, \beta_n, \delta_n)\) as “random effects” determined by an intermediate parameter \( v_n \) that has a CDF \( F(v_n | \theta) \) depending on deep parameters \( \theta \). In strategies (B) and (C), unobserved variations across consumers in tastes for product classes (e.g., “SUV’s” in automobile choice, “buses” in commuter mode choice) are introduced as additive fixed or random effects embodied in the coefficients of indicators in \( q_{jmn} \). In (C), the intermediate parameter vector \( v_n \) that determines preferences \((\alpha_n, \beta_n, \delta_n)\) can be interpreted as a hierarchical mixture given a deep parameter vector \( \theta \). In working with choice data where there is a single or limited number of observed choices for each consumer, strategy (A) is practical and often surprisingly good at predicting demand despite IIA rigidity and the failure to capture heterogeneity in \((\alpha, \beta, \delta)\), while (B) is unworkable, with no or poor identification of individual consumer taste parameters. In (C) the mixture distribution \( F(v | \theta) \) will be identified under fairly mild conditions on the distributions of attributes and the structure of utility. Alternative (C) is practical and sufficient for many purposes, can be estimated by classical or Bayesian methods, and under either estimation method can provide information at the individual consumer level that is useful in some applications, particularly when a consumer’s choice observed from a series of menus is driven by an intermediate parameter vector \( v \) drawn from a hierarchical mixing structure \( F(v | \theta) \). See Train (2009, Ch. 11, Ch. 12) for derivation of individual consumer parameters under classical estimation and under Bayesian estimation. Huber and Train (2001) describe both procedures and compare results.

The expected choice probabilities implied by the model (2) when each consumer faces a single menu \( m \) are

\[
P_k(x_{mn}, q_{jmn}, p_{mn} | \theta) \equiv \int_v P_k(x_{mn}, q_{jmn}, p_{mn} | v) \cdot F(dv | \theta) \equiv \int_v \frac{\exp \left( V_{kmn}(v) \right)}{1 + \sum_{j=1}^{J} \exp \left( V_{jmn}(v) \right)} \cdot F(dv | \theta)
\]

\[8\] A full model of intra-consumer and inter-consumer random effects could be introduced by assuming that \( F(v_n | \theta) \) describes inter-consumer heterogeneity, and another distribution \( H(v_{mn} | v_n) \) describing intra-consumer heterogeneity. It may be practical, for example, to assume that the intra-consumer variation for a positive parameter \( \alpha_m \) and unsigned parameters \( \beta_m, \delta_m \) has a symmetric distribution \( H(\log \alpha_m - \log \alpha_n, \beta_m - \beta_n, \delta_m - \delta_n) \). Note that when \( q_{jmn} \) spans all alternatives and \( \delta_{mn} \) varies across menus, it can absorb or duplicate the effect of \( \varepsilon_{jmn} \); then \( \alpha_m \) can be normalized arbitrarily at a small value. Within a hierarchical Bayesian framework, intra-consumer heterogeneity adds a hierarchical level without changing any of the structure for estimation and prediction. In terms of the three modeling strategies outlined in the first paragraph, this full model corresponds to the third strategy. The second strategy that is the focus of this paper, with homogeneous intra-consumer preferences other than “psychophysical discrimination failures” embedded in the disturbances \( \varepsilon_{jmn} \), corresponds to \( H(v_{mn} | v_n) \) specialized to unit probability at \( v_{mn} = v_n \). The first strategy, with intra-consumer heterogeneity no more extensive than inter-consumer heterogeneity, corresponds to \( F(v_n | \theta) \) specialized to unit probability at a point \( v_1 \) and \( H \) accounting for all taste heterogeneity.
a mixed (or random parameters) MNL model. A typical random parameters specification assumes an intermediate parameter vector \( \nu \) whose CDF depends on deep parameters \( \theta \), and component-wise one-to-one mappings \( \alpha(\nu), \beta(\nu), \delta(\nu) \) that are restricted in sign for taste parameters that are signed a priori. A common specification is \( F(\nu|\theta) = C(F_1(\nu_1|\theta), \ldots, F_T(\nu_T|\theta); \theta) \), where \( C \) is a copula and the \( F_i \) are the univariate marginal CDF’s of the components of \( \nu \), typically exponential, log normal, or normal with or without censoring. For example, letting \( \Phi(\nu; \Omega) \) denote a multivariate normal CDF with standard normal marginals and a correlation matrix \( \Omega \), a multivariate normal copula \( \Omega(\theta) = \Psi(\theta)^2 + \Lambda(\theta)\Lambda(\theta)' \) is the factor-analytic correlation matrix. Sampling from the multivariate normal copula with a correlation matrix \( \Omega(\theta) \) can be carried out by setting \( \eta = \Omega(\theta)^{1/2}\xi \), or in the case of a factor-analytic structure, \( \eta = \Psi(\theta)^{1/2}\xi + \Lambda(\theta)\zeta \), where \( \xi \) and \( \zeta \) are vectors of standard normal variates, and then setting \( \nu_i = F_i^{-1}(\eta_i) \) for each component.

It is also sometimes important to allow segmentation of the population, with each segment having its own distribution of \( \nu \); see Desarbo et al. (1995). Segmentation may be unobserved, with the probability of consumers falling in a segment becoming a parameter of the problem, or may be observed; e.g., classification by age or gender. An example would be a population that is segmented by preferences for “brands” or “classes” of products. The vector \( q_{jmn} \) contains indicators that are turned on for specific categories of products, and a segment might be defined as the sub-population that has positive or negative preferences for a specific class (i.e., the component of \( \delta(\nu_n) \) is non-zero for sample members \( n \) in this class, and the remaining components of \( \delta(\nu_n) \) are zero). The probability of a person falling in this sub-population would be another parameter in \( \theta \), and within each sub-population, all other taste parameters would have a (conventional) distribution conditioned on the sub-population.

### 1.3 Consumer Welfare

We next provide a method for connecting choice behavior and consumer welfare so willingness-to-pay for policy changes can be estimated; for background, see Small and Rosen (1981), McFadden

---

9 While (4) may be a parsimonious way of taking into account different levels of acuteness and whimsy across consumers, the McFadden-Train (2000) result summarized in footnote 4 shows that \( \alpha \) can be made a fixed parameter without loss of generality. Specifically, heterogeneity in acuity can be captured by a flexible specification of class identifiers \( q_{jmn} \) and class preferences \( \delta_n \) while keeping fixed the scale \( \alpha \) on the remaining error in (1). This specification may be more cumbersome than allowing \( \alpha \) to be random, but it does have an advantage later in simplifying the computation of consumer welfare.

10 A copula is a multivariate CDF \( C(u_1, \ldots, u_T) \) whose univariate marginals are all uniform \([0,1]\). Any multivariate distribution can be written in terms of its copula and marginals, and there are necessary and sufficient conditions for a function to be a copula; see Joe (1997); Fosgerau, McFadden, Bierlaire, (2013).
choice-probability-generating-function (CPGF) whose gradient is the vector of associated choice probabilities, and
give necessary and sufficient conditions for a function to be a CPGF. Their theorem in general embeds the given
random utility model in a random preference field with additive perturbations of the utilities of various alternatives,
and shows that the gradient of expected maximum utility with respect to these perturbations gives the choice
probabilities. Consider the money-metric utility function \( U(W_n,A_n,I_n,x_{jmn},q_{jmn},p_{jmn}\mid \nu_n) \) in (1) with linear additive
factors \( x_{jmn} \), \( q_{jmn} \), and \( p_{jmn} \), i.i.d. EV1 disturbances \( \epsilon_{jmn} \), and transfer income appearing solely in the linear term \( I_n \).
Playing a similar role to utility perturbations in the general case will be quality-adjusted prices \( \pi_{jmn}(v) = p_{jmn} - x_{jmn}\beta(v) - q_{jmn}\delta(v) \); these are not necessarily non-negative. Rewrite \( U \) in (1) as \( U(I_n,\pi_{jmn}\mid \nu_n) = I_n - \pi_{jmn}(\nu_n) + \alpha(\nu_n)\epsilon_{jmn} \).
Then expected maximum utility is

\[
(5) \quad G(I_n,\pi_{jmn}\mid \theta) = E_{\nu\theta} E_x \max_{j=0,\ldots,J} U(I_n,\pi_{jmn}\mid \nu) \equiv I_n + E_{\nu\theta} \alpha(\nu) \cdot \{ \log(\sum_{j=0}^{J} \exp(-\frac{\pi_{jmn}(\nu)}{\alpha(\nu)}) \} + \gamma_0 .
\]

where \( \pi_{jmn} \) denotes the vector of functions \( \pi_{jmn}(\cdot) \) and \( \gamma_0 \) denotes Euler’s constant.\(^{11}\) Then (5) has the gradient property that the negative of its derivative with respect to a price \( p_{kmn} \), or equivalently to a quality-adjusted price \( \pi_{kmn}(\cdot) \),
gives the choice probability for that alternative,

\[
(6) \quad -\partial G(I_n,\pi_{jmn}\mid \theta)/\partial p_{kmn} = P_k(x_{mn},q_{mn},p_{mn}\mid \theta) \equiv E_{\nu\theta} P_k(x_{mn},q_{mn},p_{mn}\mid \nu) \equiv E_{\nu\theta} \frac{\exp(-\pi_{kmn}(\nu)/\alpha(\nu))}{1 + \sum_{j=1}^{J} \exp(-\pi_{jmn}(\nu)/\alpha(\nu))} .
\]

The CPGF can be used to determine the effect of a policy change on consumer welfare. For a scalar \( \lambda \in [0,1] \),
let \( (x_n,^\lambda, q_n,^\lambda, p_n,^\lambda) \) denote a rectifiable path from an old policy \( (x_n,^0, q_n,^0, p_n,^0) \) to a new policy \( (x_n,^1, q_n,^1, p_n,^1) \), and let \( \pi_{jmn}^\lambda \) denote the associated quality-adjusted prices. Define compensating variation (CV\(^\lambda\)) to be the net decrease in income per capita that equates expected utility at the old policy and at \( \lambda \) along the path to the new policy, \( G(I_n - CV^\lambda, \pi_{jmn}^\lambda\mid \theta) \equiv G(I_n,\pi_{jmn}^\lambda\mid \theta) \). Let \( A \) denote the set of alternatives in \{0,\ldots,J\} that have \( \pi_{jn} \) affected by the policy change, \( B \) denote
the remaining alternatives where \( \pi_{jn}^\lambda \) is invariant, and note that \( B \) contains at least \{0\}. From (5),

\[
(7) \quad CV^\lambda = E_{\nu\theta} \alpha(\nu) \cdot \log(\frac{\sum_{j=0}^{J} \exp(-\pi_{jmn}^\lambda(\nu)/\alpha(\nu))}{\sum_{j=0}^{J} \exp(-\pi_{jmn}(\nu)/\alpha(\nu))}) \equiv E_{\nu\theta} \alpha(\nu) \cdot \log(\frac{P_B(\pi_{jmn}^\lambda\mid \nu)}{P_B(\pi_{jmn}\mid \nu)}) .
\]

Then, compensating variation \( CV^1 \) is given by the expectation with respect to stochastic taste parameters of the log
of the ratio of the before and after shares of consumers choosing alternatives not affected by the policy change.
Doug MacNair has suggested applying the expansion log(1 – y) = – y + \( O(y^2) \) to the last term in (7), giving an
approximation that is accurate when \( P_A \) is small both before and after the policy change.

\(^{11}\) The assumption of independent EV1 disturbances in (1) is not necessary for (5) to be a CPGF, but the forms (8) and (9)
for compensating variation do depend on this assumption.
The formula (8) has a particularly simple form when $\alpha$ is a constant, with $CV^1$ equal to this scale factor times the difference in the full population market share of consumers choosing the products affected by the policy change. In the formulas (7) and (8), the scaling parameter $\alpha$ appears to have a prominent role in determining the level of $CV^1$, but this is offset in the remainder of these expressions. To see this, write (7), using (6), as

\[ CV^1 = \int_0^1 \frac{dCV}{\lambda} \cdot \alpha \]  

This is the line integral over the rectifiable path of the area between the old and new quality-adjusted prices behind the demand functions for the products in $\mathbf{A}$. But this is just the classical Hicksian consumer surplus associated with this policy change when income effects are absent (e.g., no dependence of $x_n$ on $I_n$); (9) depends on $\alpha$ only through its influence on the acuity of consumer response to price changes, i.e., the price elasticities of demand. Further, these price elasticities are usually bounded even when there is a positive probability of very small $\alpha(v)$: The own price elasticity of $P_k(x_m, q_m, p_m|\theta) \cdot P_k(x_m, q_m, p_m|\theta) \cdot (1 - P_k(x_m, q_m, p_m|\theta))$. Use the inequality $e^{-c} < \alpha/c$ for $c > 0$. If $c_k = \max_{j=0}^{J} V_{jn}(v) - V_{kn}(v) > 0$, then $P_k$ is bounded by $\alpha(v)/c_k$, and if $c_k \equiv \min_{l \neq k} c_l > 0$, then $1 - P_k$ is bounded by $J\alpha(v)/c_k$. Then the price elasticity is bounded in magnitude by $\max \{p_{kmn}/c_k, j_{kmn}/c_k\}$ no matter how small $\alpha(v)$. The limited sensitivity of (7) to $\alpha$ is also seen by considering limiting cases. For constant $\alpha \to 0$, $CV^1 \to E_{\alpha,0} \{ \min_{j=0}^{J} p_{jn}^0 - \min_{j=0}^{J} \pi_{jn}^\lambda \}$, and for $\alpha \to +\infty$, $CV^1 \to E_{\alpha,0} \{ \frac{1}{J+1} \sum_{j=0}^{J} \pi_{jn}^0 - \pi_{jn}^\lambda \}$. The difference in these expressions comes only from the difference between least and average quality-adjusted prices, reflecting two extremes in the acuity of consumers in gravitating to alternatives with the least quality-adjusted price.

### 1.4 Estimation

Given data on observed choice from one menu $m$ for each consumer in a sample indexed $n = 1, \ldots, N$, the log likelihood of the observations in model (4) is

\[ \log L(\theta) = \sum_{n=1}^{N} \sum_{j=1}^{J} d_{jm} \cdot \log P_k(x_m, q_m, p_m|\theta) , \]

where $d_{jm}$ is one if consumer $n$ chooses $j$, zero otherwise. The associated score is

\[ \partial \log L(\theta) / \partial \theta = \sum_{n=1}^{N} \sum_{j=1}^{J} (d_{jm} - P_j(x_m, q_m, p_m|\theta)) \cdot \frac{\partial log P_j(x_m, q_m, p_m|\theta)}{\partial \theta} . \]

The parameters $\theta$ can be estimated by maximum likelihood, with probabilities in (5) and (6) approximated using simulation methods if necessary; see McFadden (1989, 1996), Hajivassiliou et al. (1994, 1996, 1998),
McFadden and Ruud (1994). The estimated model can then be used to simulate purchase probabilities and consumer welfare changes for counter-factual or future configurations of product attributes and prices.

The framework just described can also be applied to data on intended or stated choices obtained from hypothetical market offerings. Such data may circumvent two limitations of real market observations – lack of independent variation in some attribute dimensions and levels and in prices, due to the fact that in equilibrium markets offer only products that are sufficiently attractive to consumers to be viable, and endogeneity of product features and prices as a result of market equilibrium.\textsuperscript{12} The key questions are whether experiments with hypothetical markets and incentive structures can be designed to give context, incentives, and information similar enough to real markets to elicit the same perceptions, expectations, and cognitive processes, so that stated choice data from these hypothetical markets are predictive for choice behavior in real markets, and whether tests of the predictive validity of forecasts based on stated choice data establish a domain in which they are reliably accurate.

2. SOME HISTORY OF STATED PREFERENCE ELICITATION

Stated preference methods date back to the 1930’s, when the iconic psychologist Leon Thurstone (1931) made a presentation to the second meeting of the Econometric Society proposing direct elicitation of indifference curves. According to Moscati (2007), Thurstone’s proposal was soundly rejected by the economists of the day (e.g., Hotelling and Frisch from the floor, and Wallis and Friedman, 1942), and for almost forty years following, only demand behavior revealed in markets was considered legitimate material for mainstream economic analysis. Looking back, this seems narrow-minded, but there was some reason for it. The language of economic analysis, then and now, is prediction of market demand, and assessment of market failures in terms of dollars of equivalent lost income, deduced from demand as consumer surplus. Any measurement method that uses experimental data on preferences has to produce convincing results in this language by showing that stated preferences collected outside the market have the same predictive power for market behavior as implied preferences reconstructed from market data. With the advent of behavioral economics, we have learned that people are often not relentless utility maximizers, either in markets or in experiments, undermining the tight links neoclassical consumer theory forges between consumer utility and demand behavior. This has led to calls for less focus on market demand behavior, and assessment of market failures in terms other than dollars of lost income. This approach may eventually gain

\textsuperscript{12} Endogeneity becomes a problem for estimation of choice models when the disturbances that determine an individual consumer’s choice have components that do not “net out” at the market level. Then, equilibrium market level prices will depend on these components, and cannot be treated as predetermined in estimating the choice model parameters; see Berry-Levensohn-Pakes (1995,1998,2004). One of the attractions of choice-based conjoint analysis is that prices in offered menus can be set exogenously outside market equilibrium constraints, eliminating the endogeneity problem (provided subjects accept the offered prices as realistic and choices within the experiment are not confounded by outside market opportunities).
acceptance, but at present market prediction and valuation remain the yardsticks against which any method for eliciting consumer preferences has to be measured.

The first sustained use of stated preference methods came out of the theory of conjoint measurement pioneered by Luce and Tukey (1964) and Luce and Suppes (1965), and developed as conjoint analysis by market researchers like Paul Green (1974, 1981), Richard Johnson (1974,1999), V. Srinivasan (1974,1988,1998), and Jordan Louvier (1988) and applied to the study of consumer preferences among familiar market products (e.g., carbonated beverages, automobiles). Good introductions to conjoint experiments, data, and analysis methods are Louviere, Hensher, and Swait (2000) and Rossi, Allenby, and McCulloch (2005). A central feature of conjoint analysis is use of experimental designs that allow at least a limited mapping of the preferences of each subject, and multiple measurements that allow estimates of preferences to be tested for consistency and refined when necessary. Early conjoint analysis experiments described hypothetical products in terms of price and levels of attributes in various dimensions, and asked subjects to rank attributes in importance, and rate attribute levels and products. These measurements were used by market researchers to classify and segment buyers, and target advertising, but they were not reliable tools for predicting market demand. However, McFadden et al. (1986) and McFadden (1986) showed how choice-based conjoint elicitations could be analyzed using the tools of discrete choice analysis. Subjects would be presented with a series of menus of products. Each product offered in each menu would be described in terms of price and levels of attributes. Subjects would be asked to choose their most preferred product in each menu. For example, subjects would be offered a menu of paper towels, with each product described in terms of price, number of towel sheets, a measure of the absorption capacity, a measure of strength when wet, and brand name. Choice data from these menus, within and across subjects, could then be handled in the same way as the real market choice data described in Section 1. These choice-based conjoint (CBC) analysis methods have become widely used and accepted in market research to predict the demand for consumer products, with a sufficient track record so that it is possible to identify some of the necessary conditions for successful prediction; see Green et al. (2001), Cameron et al. (2013), McFadden (2014b).

Starting in the 1950’s, environmental economists developed a simple stated preference method termed contingent valuation (CV), and applied it to valuing environmental damage. Its development has occurred almost entirely within a tight circle of environmental economists who emphasize the unique features of environmental applications and have been selective in incorporating findings from research in marketing, cognitive psychology, and behavioral economics; see Carson, Flores, and Meade (2001). CV can be viewed as a truncated form of conjoint analysis with two important differences. First, it does not have the full experimental design features of conjoint analysis that allow tests for the structure and consistency of stated preferences. Second, CV as it has developed has been utilized primarily for valuation of environmental public resources, and as a consequence usually does not have
predictive accuracy as a direct yardstick for reliability. Instead, it relies indirectly on internal consistency checks, and on analogies with stated preference studies of consumer goods in markets and the tight links between preferences, demands, and valuations that hold when neoclassical consumer theory is valid.

Other elicitation methods for stated preferences, termed vignette analysis and measurement of subjective well-being, have become popular among some applied economists and political scientists; see Rossi (1979), King et al. (2004), Caro et al. (2012), Kahneman and Krueger (2013). The first uses detailed descriptions of alternatives, often visual, and may improve consumer information and understanding. The second elicits overall self-assessments of welfare. These self-assessments resemble the ratings collected in the early days of conjoint analysis, and may be similarly problematic for predicting choice behavior.

Since the focus of this paper is market demand forecasting for new or modified products, we do not attempt here any overall assessment of the reliability of contingent valuation, vignette analysis, or subjective well-being methods in their primary uses. We caution that in the instances where these methods have been tested, they have proven to be fairly strongly influenced by context and anchoring effects that in market demand forecasting applications may reduce forecast accuracy; see Carson (2012), Hausman (2012), Kling et al. (2012) for discussions of the particular challenges of use of contingent valuation for natural resource valuation.

3. CHOICE-BASED CONJOINT ANALYSIS (CBC)

A choice-based conjoint analysis (CBC) offers a consumer a series of menus of alternative products with profiles giving levels of their attributes, and asks him or her to identify which product he/she most prefers in each menu. The menus of products and their descriptions are designed to realistically mimic a market experience, where a consumer is presented with various competing alternatives and chooses one of the options. By changing the attribute levels available for the included products and presenting each consumer with several menus, the researcher obtains information on the relative importance that the consumer places on each of the attributes. The classic CBC setup in marketing might be a laboratory experiment where subjects are asked to sample actual products with the different profiles, and then asked for their choices from different menus. For example, subjects might be given tastes of cola drinks from menus, with various degrees of sweetness, carbonation, flavor, and price for the different products, and asked to pick one or none from each menu. However, CBC can also be used for familiar products whose features are simply described in words and pictures, with subjects asked to choose from a menu of products based on these descriptions. For example, experiments on automobile brand and model choice describe alternatives in terms of price and attributes such as horsepower, fuel consumption, number of seats, and cargo space. These studies can determine with considerable predictive accuracy the distributions of preference weights that consumers give to various vehicle features, and the automobiles they will buy; see Urban et al. (1990, 1997); Brownstone and Train (1999); Train and Winston (2007). The levels of attributes of the products offered on different menus can be
set by experimental design so that it is possible to separate statistically the weights that consumers give to the
different attributes. In its early days, menu designs were often of a “complete profile” form that mimicked classical
experimental design and allowed simple computation of “part-worths” from rating responses, but currently the
emphasis is simply on ensuring that menus are realistic and incorporate sufficient independent variation in the
attributes so that the impact of each attribute on choice can be isolated statistically.

Conjoint analysis methods can be expected to work relatively well for preferences among consumer market
goods when the task is choice among a small number of realistic, relatively familiar, and fully described alternatives,
with clear and well-understood incentives for truthful response. The idea behind incentives is that when subjects
have a realistic chance of really getting what they say they prefer, and they understand this, they have a positive
disincentive to misrepresent their preferences and risk getting an inferior outcome. Studies of conjoint methods
show that they are in general less reliable and less directly useful for predicting behavior when the task is to rate
products on some scale, or to adjust some attribute (e.g., price) to make alternatives indifferent; these seem to induce
cognitive “task-solving” responses different from the task of maximizing preferences. Asking follow up questions
about a single menu also seems to induce a different mind-set than simple choice. For example, a study might
follow up a stated choice with a question about the second best choice among the remaining alternatives, a question
as to whether the consumer would stay with first stated choice if the price of one of the alternatives were reduced,
or questions about the perceived attributes of various alternatives. Empirical experience is that such follow-up
questions elicit responses that are not always consistent with the initial stated choices, even though they do not
differ much in their framing from market experiences. The explanation may be that the initial menu “anchors”
perceptions and shadows subsequent responses, and that follow-up questions induce a “bargaining” mind-set that
invites strategic responses; see Hanemann-Loomis-Kanninen (1991), Green et al. (1998). Follow-up questions on
perceptions can also be colored by self-justification. Conjoint methods can be expected to be less reliable when the
products are unfamiliar or incompletely described, or involve public good aspects that induce respondents to
incorporate social welfare judgments; e.g., when preferences for automobile models are stated in an elicitation that
emphasizes the energy footprint of the models, and environmental consequences. Valuation of non-use aspects
of natural resources are particularly challenging for conjoint methods because these applications seek to measure
preferences that are outside normal market experience of consumers.

13 See Wright and Kriewall (1980); Chapman and Staelin (1982); Elrod et al. (1992).

14 This is not an argument that energy footprint should be excluded from the attribute list in studying automobile choice,
rather a caution that when it is included, responses are likely to be quite sensitive to how statements about energy footprints
and environmental consequences are framed.
We discuss six important issues that need to be considered when designing a CBC study. *First*, familiarity is important. If subjects are experienced with the products or services, and the attributes that are being assessed, then they seem to make more consistent and predictive choices. If possible, subjects should have the opportunity to test for themselves their response to different attributes. For example, in a study of consumer choice among streaming music services with various attributes, it should improve prediction to give subjects hands-on experience with different features that is similar to their opportunities to investigate and experience these features in a real market. This might be done with mock-up working models of the products, or with computer simulation of their operation. However, there is a tradeoff: attempting to train consumers, and providing mock-ups, can inadvertently create anchoring effects. Consumers who are unfamiliar with a product may take the wording in the training exercises about the attributes, and the characteristics of the mockups, as clues to what they should feel about each attribute. Even the mention of an attribute can give it more prominence in a subject’s mind than it would have otherwise. The researcher needs to seriously consider the sometimes-conflicting goals of making the subject knowledgeable about the products without influencing their relative values of attributes. *Second*, the researcher needs to decide whether to offer an “outside alternative” in the choice sets, and, if so, how to characterize it to the subjects. The inclusion of a realistic “no purchase” option allows estimation of market shares and price elasticities, while experiments without this option can only be used to estimate demand conditional on a purchase. However, if the outside option is included, it is important that the meaning of the option be clearly delineated to subjects. E.g., in a car choice exercise, does “no purchase” mean that the subject would use a vehicle that the household currently owns and reconsider options next year, or what? The danger is that the “no purchase option” can be interpreted differently by different subjects, and can easily become a way for subjects avoid the effort of resolving difficult tradeoffs. Whether and how to include an outside option is an important decision that the researcher must make. If it is not included, it may be necessary to use external market share data to constrain or calibrate the choice model fitted to the CBC data so that it can make complete market demand predictions. *Third*, if possible, the conjoint study should be incentive compatible, so that subjects have a positive incentive to be truthful in their responses. For example, suppose subjects are promised an Visa® cash card, and then offered menus and asked to state whether or not they would purchase a product with a profile of attributes and price, with the instruction that at the end of the experiment, one of their choices from one of their menus will be delivered, and the price of the product in that menu deducted from their cash card balance. If they never choose the product, then they get no product and the full Visa balance. These subjects learn, perhaps with training or through experience, that it is in their interest to say they would choose a product if and only if its value to them is higher than its price. Then, they have a positive incentive to be truthful, and the experiment is said to be incentive-aligned or incentive-compatible. In many situations it will not be practical to provide an incentive compatible format while maintaining the objectives of the analysis. For example, the researcher might want to consider combinations of attributes that cannot be constructed at this time.
Or the cost of providing a product at any meaningful probability is prohibitive. For example, automobile manufacturers might want to test consumers’ reactions to new features during the design phase of the manufacturers’ research. And for existing attributes, it is not really feasible to offer subjects enough money to buy a new car and then provide them one of the chosen cars at the listed price. Creating the incentive compatibility probabilistically might require so low a probability to be practical that the subjects do not take the offer seriously (e.g., a one in a thousand chance of being offered the vehicle you chose in one menu at its stated price.) The researcher should ensure incentive compatibility when possible, realizing, however, that CBC experiments without incentive compatibility have been found to use useful in many settings; see Dong et al. (2010). Fourth, the researcher needs to decide how “far down” to explore stated preference orderings. Subjects’ first choice (i.e., most preferred option) is most natural to consumers, since it mimics their regular purchasing task. Second choice, third choice, etc., can be colored by framing dynamics and may be less reliable for predicting market behavior; see McFadden (1981), Green et al. (1998), and Louviere’s (1988) “best-worst” choice setup. Fifth, where possible, CBC results should be tested against and calibrated to consumer behavior in real markets. In some cases, CBC menus will coincide with product offerings in existing markets. In this case, it is useful to compare models estimated from the CBC study and the market data to assess whether people are weighing attributes similarly. Improved forecasts may be obtained by imposing real market constraints such as product shares on the estimation of choice models from CBC data, by calibrating CBC model parameters to satisfy market constraints, or by combining CBC and market choice data and estimating a combined model with scaling and shift parameters for CBC data as needed. Sixth, CBC studies should when possible embed tests for response distortions that are commonly observed in cognitive experiments, such as anchoring to cues in the elicitation format, reference point or status quo bias, extension neglect, hypersensitivity to context, and shadowing from earlier questions and elicitations. While some of these cognitive effects also appear to influence market choices, many are specific to the CBC experience and have the potential to reduce forecasting accuracy. Ideally, a well-designed CBC study will not show much sensitivity of its bottom-line WTP values to these sources of possible response distortion.

The following sub-sections discuss in more detail these conditions for reliable demand prediction using stated preference data.

3.1. Sampling and Recruitment: Target populations may differ depending on the objectives of the study – e.g., current users, current and potential users, the general population. It is important that the sampling frame draw randomly from the target population, without excessive weighting to correct for stratification and non-response. However, not all members of the target population may have the background needed to make informed product choices. Then it may be more informative to study the preferences of experienced users, and separately study the differences in users and non-users. An example might be study of consumer demand for relatively esoteric technical
attributes of products, say the levels of encryption built into telecommunications devices, where only technically savvy device users will appreciate the meaning of different encryption levels. In this case, a good study design may be to conduct an intensive conjoint analysis on technically knowledgeable users, and separately survey the target population to estimate the extent and depth of technical knowledge, and the impact of technical information on the purchase propensities of general users and non-users.

Relatively universal internet access has led to less expensive and more effective surveying via the internet than by telephone, mail, or personal interview. However, it is dangerous in internet surveys to recruit subjects by advertising for and sampling from volunteers, as they are likely to look quite different overall than a target population of possible product buyers. Better practice is to recruit subjects using a reliable method such as random sampling of addresses, then recruitment of subjects from the sampled addresses. It is important to compensate subjects for participation at sufficient levels to minimize selection due to attrition. Experience with “professional” subjects who are paid to participate in internet panels is positive: subjects who view responding as a continuing “job” with rewards for effort are more attentive and responsive. In a study of mechanisms for social choice, McFadden (2012) finds the overall social cost of survey responses that lead poor choices is minimized with large participation fees (over $100) that induce non-participation rates below 10 percent.

3.2. Design of Conjoint Experiments. The design of a conjoint experiment establishes the number of menus offered to each subject, the number of products on each menu, the number of attributes and attribute levels introduced for each product, and the design of the profiles of the products placed on each menu. Some other aspects of a conjoint study, the setup and introduction to the experiment given to each subject, subject training, and incentives, might be considered components of the design, but will here be treated separately. There are three distinct considerations that enter conjoint experimental designs.

The first consideration is that for good statistical identification of the valuations of separate attributes, the design needs to allow considerable independent variation in the levels of different attributes, and a considerable span of attribute levels. The classical statistical literature on experimental design focused on analysis of variance and emphasized orthogonality properties that permitted simple computation of effects, and treatments that provided minimum variance estimates. Designs that reduce some measure of the sampling variance under specified model parameters (such as the determinant of the covariance matrix for “D-efficiency”) have been implemented in market research by Kuhfield et al. (1994), Bleimer and Rose (2009), Rose and Bliemer (2009),and others. It is important that conjoint studies be designed to yield good statistical estimates, but there is relatively little to be gained from adherence to designs with classical completeness and orthogonality properties. First, with contemporary computers, the computational simplifications from orthogonal designs are usually unimportant. Second, for the nonlinear models used with CBC, orthogonality of attributes does not in general minimize sampling variance. Unlike classical
analysis of variance problems, it is not usually possible in nonlinear choice models to specify in advance designs that are efficient without knowing the parameters that are the target of the analysis.

The second consideration is that relatively mechanistic statistical approaches to setting attribute levels may lead to profiles that are unrealistic, or are dominated by the profiles of other products on a menu. Considerable care is needed to balance statistical objectives with realism of the experiment; see Huber and Zwerina (1996). Menus and their framing that are unlike familiar market settings invite cognitive responses that differ from those that appear to determine preferences and drive choices in market settings. There is a tendency for subjects to approach surveys as if they were school exams – they cast about for “correct” answers by making inferences on what the experimenter is looking for. While some may use their responses to air opinions, most give honest answers, but not necessarily to the question posed by the experimenter. They may “solve” problems other than recovering and stating their true preferences, indicating instead the alternative that seems the most familiar, the most feasible, the least cost, the best bargain, or the most socially responsible; see Schkade and Payne (1993).

The third consideration is that prominence and ease of comparison are known to be factors that influence the attention subjects give to different aspects of decision problems; for example, there is a claim that subjects in their stated choices systematically place more weight on price relative to other product attributes than they do in real markets, perhaps because this dimension is clearly visible and comparisons are easy in a conjoint analysis menu, whereas prices in real markets often come with qualifications and often are not displayed side-by-side. Widespread folklore in marketing is that subjects have trouble processing more than six attributes and more than four or five products, and begin to exhibit fatigue when making choices from more than twenty menus; see Johnson and Orme (1996). Beyond these limits, they appear to take “short cuts” by eliminating consideration of some products and attributes using simple heuristics, and considering tradeoffs only on the remainder. Often conjoint analysts will address this behavior by limiting the dimensionality of the attribute profile, explicitly or implicitly asking subjects to assume that in all other dimensions, the products are comparable to brands currently in the market. This leaves subjects free to make possibly heterogeneous and unrealistic assumptions about these omitted attributes, or requires them to digest and remember lengthy specifications for omitted attributes and their assumed levels. These design restrictions may make responses more consistent and easy to analyze, but they may not improve prediction. Filtering heuristics also seem to be used in real markets with many complex products such as the market for houses. If the primary focus of the conjoint study is prediction, then the best design may be to make the experiment as realistic as possible, with approximately the same numbers and complexity of products as in a real market with the same line, so that consumers face similar cognitive challenges and respond similarly even if decision-making is less single-minded than neoclassical preference maximization. However, if the primary focus is measurement of consumer welfare, there are deeper problems in linking well-being to demand behavior that reflects filtering. While
it may be possible to design simple choice experiments that eliminate filtering and give internally consistent
statements of well-being, there is currently no good theoretical or empirical framework for aligning such
measurements with neoclassical economic welfare analysis or validating them using real market choice data that
reflects filtering.

3.3. Subject Training. Extensive experiments from cognitive psychology show that context, framing, and
subject preparation can have large, even outsize, effects on subject response. It is particularly important that subjects
have familiarity with the products and features they are being asked to evaluate that is comparable to their real
market experiences, as context and framing effects are particularly strong when subjects are asked to respond in
unusual or unfamiliar circumstances. Familiarity may be automatic if the target population is experienced users of
a particular line of products. For inexperienced users, tutorials on the products and hands-on experience can reduce
careless or distorted responses, but also increase the chance of influencing stated preferences in ways that reduce
forecasting accuracy.

It is useful to recognize that training of subjects can occur at several levels, and that training can manipulate
as well as educate, leading to unreliable demand predictions. First, subjects have to get used to answering questions
that may be difficult or intrusive, and learn that it is easier or more rewarding to be truthful than to fabricate. Some
of this is mechanical: practice with using a computer for an internet-based conjoint survey, and moving through
screens, buttons, and branches in a survey instrument. Second, subjects need to be educated as to what the task of
stating preferences is. Subjects can be taught in “Decision-Making 101” how to optimize outcomes with assigned
preferences, and how to avoid mistakes such as confusing the intrinsic desirability of a product with process issues
such as availability or dominance by alternatives. Such training can be highly manipulative, leading to behavior
that is very different from and not predictive for real market choices. But real markets are also highly manipulative,
providing the “street” version of “Decision Making 101”, teaching by experience the consequences of poor choices.
The goal of a conjoint study designed for prediction should be to anticipate and mimic the training that real markets
provide. Third, the study designer needs to determine what information will be conveyed to the subject, in what
format, and assess what information the subject retains and understands. Typically a conjoint survey begins by
describing the types of products the subject will be asked to evaluate, their major attributes, and the structure of the
elicitation, asking for most preferred alternatives from a series of menus. Details may be given on the nature and
relevance of particular attributes. Instructions may be given on the time the subject has to respond, and what rules
they should follow in answering. For example, the survey may either encourage or discourage the subject from
consulting with other family members, finding and operating past products in the same line, or consulting outside
sources of information such as internet searches. Finally, subjects need to be instructed on the incentives they face,
and the consequences of their stated choices. At various stages in subject training, they may be monitored or tested
to determine if they have acquired information and understand it. For example, a protocol in market research called
*information acceleration* gives subjects the opportunity to interactively access product descriptions, consumer
reviews, and media reports, and through click-stream recording and inquiries during the choice process collects data
on time spent viewing information sources and impact on interim propensities. This protocol seems to improve
subject attention and understanding of product features, and also identify the sources and content of information
that has high impact on stated choices; see Urban et al. (1990,1997).

In summary, while training may educate subjects until they are familiar with the products being compared, it
is difficult to design training that is neutral and non-manipulative. Note that real markets are in fact often
manipulative, via advertising and peer advice, and one goal for CBC is to achieve accurate prediction by mimicking
in the experiment the advertising and other elements of persuasion the consumer will encounter in the real market.
One caution is that particularly in cases where preferences are not well-formed in advance, subjects will be
particularly vulnerable to manipulation, and training that embodies manipulation that is not realistic risks inducing
stated responses that are not predictive for real market behavior.

3.4. Incentive Alignment. Economic theorists have developed a number of mechanisms for incentive-
compatible elicitation of preferences. The simplest offer the subject a positive probability that every stated choice
will result in a real transaction. If subjects understand the offer, the probabilities are sufficiently high so that the
subject does not dismiss them as negligible\(^\text{15}\), and subjects view the transactions as being paid for from their own
budgets rather than in terms of “house money” that they feel is not really theirs, then it is a dominant strategy for
the subject to honestly state whether or not a product with a given profile of attributes is worth more than its
price; this is shown by Green et al. (1998) for a leading variant of this mechanism due to Becker, DeGroot, Marschak
(1964). More complicated mechanisms have been shown to be incentive-compatible in circumstances where
choices involve social judgments as well as individual preferences; for example, McFadden (2012) shows how the
Clark-Groves mechanism can be used in an economic jury drawn at random from the population to decide on public
projects. However, experimental evidence is that people have difficulty understanding and accepting the incentives
in these mechanisms, and acting on them. Thus, it can be quite difficult in practice to ensure incentive alignment
in a CBC, or to determine in the absence of strong incentives whether subjects are responding truthfully.

\(^\text{15}\) Suppose a CBC experiment on choice among automobiles offers a one in ten-thousand chance of receiving either a
$40,000 car or the cost of the car in cash, depending on stated choice. If the true value of the car to this subject is $V, then a
truthful response is to choose the car if and only if $V > $40,000. If a consumer declines the car when $V > $40,000, then his
expected loss is $V/10000 - $4, a small number. This incentive is still enough in principle to induce the rational consumer to
state truthfully whether or not he prefers the car. However, misperceptions of low-probability events and attitudes toward risk
may in practice lead the consumer to ignore this incentive or view it as insufficient to overcome other motivations for
misleading statements.
Fortunately, there is also considerable evidence that while it is important to get subjects to pay attention and answer carefully, they are mostly honest in their responses irrespective of the incentives offered or how well they are understood; see Bohm (1972), Bohm et al. (1997), Camerer and Hogarth (1999), Yadav (2007), Dong et al. (2010).

3.5. Reconciliation and Validation. An advantage of CBC experimental designs is that through the presentation of a slate of menus, they give an opportunity to test the consistency of individual stated choices with neoclassical preference theory, to confront respondents and ask them to explain and reconcile stated choices, and to incorporate menus that allow direct cross-validation between stated and revealed market choices. For example, menus can be offered that allow for the possibility of testing whether stated choices are consistent with the axioms of revealed preference, and specifically whether they violate the transitivity property of preferences. Even under the more relaxed standard that consumers have stochastic preferences with new preference draws for each choice, their responses can be tested for the regularity property that adding alternatives cannot increase the probability that a previous alternative is chosen. If menus contain current market alternatives, and past purchase behavior of the subjects is known, then one can test whether revealed and stated preferences for the same alternatives are consistent in their weighting of attributes. For example, Morikawa, Ben-Akiva, McFadden (2002) find that there are systematic differences in weights given to attributes between stated and revealed choices, and that predictions from stated choices can be sharpened by calibrating them to revealed preferences; see also Ben-Akiva and Morikawa (1990), Brownstone et al. (2000), Hensher and Bradley (1993). This step of testing and validating CBC is important particularly in studies where verisimilitude of the conjoint menus and congruity of the cognitive tools respondents use in experimental and real situations are in question, for example when the products being studied are complex and unfamiliar, such as choices of college, house to purchase, cancer treatment to pursue, or remedies for environmental damages.16

3.6. A Conjoint Study Example. To make CBC more concrete, consider the problem of estimating consumer preferences for table grapes, to guide planting and pricing decisions. In preliminary focus groups, it is found that in addition to price per pound, the attributes that consumers mention as important are sweetness, crispness, size, and whether or not they are organically grown. Subjects interviewed at a mall are given eight menus, with each menu containing the no-buy alternative and three alternative types of grapes pictured and described in words giving their price per pound and attribute levels. The CBC experimental design varies price and product attributes over various levels, and assigns these randomly to bunches in each menu, making sure that the bunches offered all differ

from each other in two or more dimensions. The levels in the study are $1.00 to $4.00 for Price, {Tart, Sweet} for Sweetness, {Soft, Crisp} for Crispness, {Small, Large} for Size, and {No, Yes} for Organic. These “raw” attributes of the offered alternatives enter the utility model (1); this model may also include transformations of these attributes, such as an interaction of the sweetness and crispness attributes, (Sweetness = 1)*(Crispness = 1) or an interaction of a product attribute such as sweetness and subject gender, (gender==female)*(sweetness = 1). Subjects are told that at the end of the interview, one of the menus they have responded to will be drawn at random, and they will receive their choice from this menu plus a Visa cash card for $10 less the price of their choice. This CBC design has strong incentive alignment, provided subjects believe that the promised transaction will be carried through. Table 1 gives an example of a menu:

<table>
<thead>
<tr>
<th>Menu 1</th>
<th>Bunch #1</th>
<th>Bunch #2</th>
<th>Bunch #3</th>
<th>No Purchase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>$2.50</td>
<td>$2.75</td>
<td>$3.00</td>
<td>$0.00</td>
</tr>
<tr>
<td>Sweetness</td>
<td>Tart (0)</td>
<td>Sweet (1)</td>
<td>Sweet (1)</td>
<td></td>
</tr>
<tr>
<td>Crispness</td>
<td>Crisp (1)</td>
<td>Soft (0)</td>
<td>Crisp (1)</td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>Small (0)</td>
<td>Large (1)</td>
<td>Small (0)</td>
<td></td>
</tr>
<tr>
<td>Organic?</td>
<td>No (0)</td>
<td>No (0)</td>
<td>Yes (1)</td>
<td></td>
</tr>
<tr>
<td>Your Choice</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

This example and its incentive scheme show some of the problems in designing a conjoint survey to closely mimic the real market. If some of the attribute combinations do not occur in existing products, then it may not be possible to fulfill a promised transaction. Incentives can be aligned through more general promises to deliver an available product (or cash) consistent with the subject’s stated preferences, but it may be more difficult for the subject to recognize that this makes it in their interest to be truthful. The subjects are not given grapes to taste that define sweetness or crispness, so must rely on familiarity and experience with table grapes to give terms like “tart” and “soft” meaning. Not all subjects will have the same interpretation, so there is a problem of tying stated preferences along these dimensions to demand for specific products that may be introduced into the real market in the future. Whether subjects view the offered prices as reasonable, or grapes as tempting, will depend on their unknown shopping histories, particularly response to promotions and sales, their anticipations of how they will feel if they receive grapes, and their expectations regarding availability and prices of grapes elsewhere. In particular, it is important to clarify the meaning of the “no purchase” option. Without direction, subjects will probably think of the offered menus in the context of their past consumption and current stock of grapes at home, and of their options for purchasing grapes when they next go to the food market. These factors also enter real market purchase decisions,
so this context may be realistic, but without measurement or control it is risky to assume this. Alternately, subjects could be instructed to assume that they have no grapes at home, and no opportunity outside the experiment to purchase grapes today. The question in this case becomes whether they understand these instructions and accept them. Overall, the usefulness of having the “no purchase” option will depend on its being given a sufficiently specific description and context so that it corresponds realistically to the options the consumer will have in the forecast market.

Table 2. Extract of Data from the Table Grape CBC

<table>
<thead>
<tr>
<th>Subject</th>
<th>Menu</th>
<th>Alternative</th>
<th>Price</th>
<th>Sweetness</th>
<th>Crispness</th>
<th>Size</th>
<th>Organic?</th>
<th>Sweet*Crisp</th>
<th>S*Gender</th>
<th>Bunch intercept</th>
<th>Gender</th>
<th>Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2.98</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1.51</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1.42</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2 illustrates the data obtained from this CBC, with the attributes and levels coded as indicated in Table 1. In addition, “Bunch intercept” is a dummy variable that is one except for the no–purchase alternative, and “Choice” is the choice indicator that appears in (8) for subject n. The data continues with the remaining menus for the first subject, next the data for the second subject, and so on.

3.7. Stated Perceptions, Expectations, and Well-Being. The focus of CBC analysis as it has developed in the literature is on preferences among products. In principle, recovery of preferences can incorporate subjective probabilities through study of products whose delivery is contingent on verifiable future events. As developed by Ramsey and Savage, and outlined by Arrow (1971), it is necessary only that markets or experiments offer a rich family of lotteries that differ only in the events on which they pay off, that consumers have complete, transitive, continuous preferences over contingent goods and services, and that consumers satisfy some plausible axioms on the separation of subjective probabilities of events and preferences over consequences of events. However, it has been recognized since the work of Allais (1953), and emphasized by the experiments of Kahneman and Tversky (1979,2000), that consumers are sensitive to context when asked to perform cognitive tasks requiring personal
probabilities. Consequently, the task of recovering consistent, stable personal probabilities is at least as challenging as the task of recovering preferences. At this point, there is no systematic approach to measuring perceptions that is comparable to CBC analysis of perceptions, although one can in principle conduct CBC studies of choice among contingent events. Manski (2004, 2010) has found that stated probabilities on a 100-point scale can be relatively consistent and fairly predictive. The Kahneman-Tversky experiments indicate however that stochastic elements and systematic biases are likely to be intrinsic to subjective probability judgments, and a problem for the future is to combine Manski-type elicitations, a framework for incentive alignment, and experimental designs like those of contingent analysis, to map out models of personal probabilities and risk response that are predictive for market demand for products that have contingent attributes.

4. MAXIMUM SIMULATED LIKELIHOOD (MSL) ANALYSIS OF CBC DATA

The CBC elicitation format produces data on choices from hypothetical market experiments which must then be analyzed to model preferences. Simulations from these preferences can then be used to predict market demands for products in the future with different attribute profiles or prices. A simple approach to these tasks is to treat the choice data from a CBC study exactly like observed market choice data, without accounting for statistical dependence of stated choices from a slate of menus by each consumer. If a flat MNL model is believed to be sufficient to explain and forecast real market choices, without introducing heterogeneity in preference weights for different product attributes, then it may also suffice for this application. Suppose data has been collected on \( n = 1, \ldots, N \) subjects, each of whom has been presented \( m = 1, \ldots, M \) menus containing \( j = 1, \ldots, J \) products and the no-purchase option \( j = 0 \). Let \( x_{jm} \) denote the attribute profile of product \( j \) on menu \( m \) for subject \( n \), \( q_{jm} \) denote a vector of dummy variables that reflect the “class” structure of alternatives, and \( p_{jm} \) denote price. Let \( x_n = (x_{1mn}, \ldots, x_{Jmn}) \), \( q_n = (q_{1mn}, \ldots, q_{Jmn}) \), and \( p_n = (p_{1mn}, \ldots, p_{Jmn}) \) denote vectors of these profiles and prices. Then, the flat MNL log likelihood of the stated choice data is

\[
\log L(\alpha, \beta, \delta) = \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{j=0}^{J} d_{jm} \log \left( \frac{\exp(x_{jm} \beta + q_{jm} \delta - p_{jm})}{\alpha + \sum_{i=1}^{J} \exp(x_{im} \beta + q_{im} \delta - p_{im})} \right),
\]

where \( d_{jm} \) is an indicator for the alternative chosen on menu \( m \) for subject \( n \).

Now allow for preference heterogeneity across consumers. As discussed earlier, CBC studies normally cannot make the number of menus \( M \) large enough to make estimation of “fixed effects” taste coefficients for each consumer reliable. The only practical alternative in most cases is to make \( \alpha(\nu), \beta(\nu), \delta(\nu) \) “random effects” drawn for each consumer from a parsimoniously parameterized distribution described by a density \( f(\nu|\theta) \) that depends on deep parameters \( \theta \). For simulation, it is useful to also write \( \nu \) as a transformation \( \nu = V(\zeta, \theta) \) of a random vector \( \zeta \).
that has a “standard” density $\varphi(\zeta)$. When the parameters for a consumer are homogeneous over the menus $m = 1, \ldots, M$, so that the consumer is a neoclassical utility maximizer except for the stochastic components $\alpha(v)\epsilon_{jmn}$ that reflect psychometric noise and tie-breaking rules, the likelihood for this consumer is

$$L_n(d_n|x_n, q_n, p_n, \theta) = \int f(\nu|\theta) d\nu \equiv \int \varphi(\zeta) d\zeta.$$  

where $P_n(\nu) \equiv \prod_{m=1}^{M} \prod_{j=0}^{J} \exp\left( \frac{V_jmn(\nu)}{\alpha(\nu)} \right) \sum \exp\left( \frac{V_imn(\nu)}{\alpha(\nu)} \right)$. The gradient of (13) is

$$\frac{\partial L_n(d_n|x_n, q_n, p_n, \theta)}{\partial \theta} = \int f(\nu|\theta) d\nu \equiv \int \varphi(\zeta) d\zeta.$$  

Standard maximum likelihood theory states that under mild regularity conditions that will ordinarily be satisfied in this application, iteration of the BHHH condition

$$\sqrt{N}(\theta^{i+1} - \theta^i) = (\Sigma_N)^{-1} \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \left( \frac{\partial \log L_n(d_n|x_n, q_n, p_n, \theta^i)}{\partial \theta} \right),$$  

where

$$\Sigma_N = \frac{1}{N} \sum_{n=1}^{N} \left( \frac{\partial \log L_n(d_n|x_n, q_n, p_n, \theta^i)}{\partial \theta} \right) \left( \frac{\partial \log L_n(d_n|x_n, q_n, p_n, \theta^i)}{\partial \theta'} \right)^{\prime},$$  

with an added step-size adjustment to avoid overshooting and accelerate convergence, is guaranteed to converge from any initially consistent estimator $\theta^0$ to an estimator $\hat{\theta}$ that maximizes the likelihood of the sample $L(\theta) \equiv \prod_{n=1}^{N} L_n(d_n|x_n, q_n, p_n, \theta)$, and further that $\sqrt{N}(\hat{\theta} - \theta^0)$ is asymptotically normal with mean zero and a covariance matrix that is consistently estimated by $(\Sigma_N)^{-1}$. Note that this result does not guarantee convergence to the maximum likelihood estimator from all starting values for $\theta$, nor can one in general be assured of the global strict concavity that would imply such global convergence. Consequently, it is important to start the MLE iteration from a grid of starting values, and perhaps introduce annealing (discussed in the next section) during the iterations, to assure that search spans the parameter space and has a high probability of finding a global rather than a local maximum.

---

17 Berndt et al. (1974). Alternately, Gauss-Newton or Newton-Raphson iteration to the maximum can be used.
Maximum simulated likelihood (MSL) replaces $L_n(d_n | x_n, q_n, p_n, \theta) = \int \mathcal{P}_n(V(\zeta, \theta)) \cdot \varphi(\zeta) d\zeta$ with the approximation $\frac{1}{R} \sum_{r=1}^{R} \mathcal{P}_n(V(\zeta_r, \theta))$, where the $\zeta_r$ are $R$ independent pseudo-random draws from the density $\varphi(\zeta)$, and uses a commensurate approximation $\frac{1}{R} \sum_{r=1}^{R} \frac{\partial \mathcal{P}_n(V(\zeta_r, \theta))}{\partial \theta}$ to the gradient $\frac{\partial L_n(d_n | x_n, q_n, p_n, \theta)}{\partial \theta}$ to maximize the simulated likelihood using (15); see McFadden and Ruud (1994). Alternately, a method of simulated score (MSS) uses (15) to find a root of the score of the likelihood, with $L_n(d_n | x_n, q_n, p_n, \theta)$ and $\frac{\partial L_n(d_n | x_n, q_n, p_n, \theta)}{\partial \theta}$ approximated respectively by $\frac{1}{R} \sum_{r=1}^{R} \mathcal{P}_n(\nu_r)$ and $\frac{1}{R} \sum_{r=1}^{R} \text{log } f(\nu | \theta)$, where the $\nu_r$ are independent pseudo-random draws from $f(\nu | \theta)$; see Hajivassiliou and McFadden (1998). If $R$ rises at least as rapidly as $\sqrt{MN}$, then either of these simulation estimators will be asymptotically equivalent to the (intractable) exact maximum likelihood estimator. Train (2007) points out that the gradient approximations in MSL and MSS are not identical for fixed $R$, even when generated by the same random draws, and that for numerical stability, MSL should use its commensurate gradient approximation.

Suppose $F(\nu | \theta)$ is multivariate normal with density $n(\nu; \mu, \Omega)$, where $\Omega = \Lambda \Lambda'$ and $\theta = (\mu, \Lambda)$, and $\alpha(\nu), \beta(\nu), \delta(\nu)$ are known component-wise transformations of $\nu$. Then, $\log n(\nu; \mu, \Omega) \propto -\log | \Lambda | - (\nu - \mu)^T(\Lambda \Lambda')^{-1}(\nu - \mu)/2$ implies $\frac{\partial \log n(\nu; \mu, \Lambda \Lambda')}{\partial \mu} = \Omega^{-1}(\nu - \mu)$ and $\frac{\partial \log n(\nu; \mu, \Lambda \Lambda')}{\partial \Lambda} = \{\Omega^{-1}(\nu - \mu)(\nu - \mu)^T \Omega^{-1} - \Omega^{-1}\} \Lambda$, and the gradient of (13), given by Train (2007), is:

$$\frac{\partial \frac{L_n}{\partial \mu}}{\partial \frac{\partial \mu}{\partial d_n}(d_n | x_n, q_n, p_n, \theta) = \int \mathcal{P}_n(\nu) \cdot \Omega^{-1}(\nu - \mu) \cdot F(d\nu | \theta)$$

$$\frac{\partial \frac{L_n}{\partial \Lambda}}{\partial \frac{\partial \Lambda}{\partial d_n}(d_n | x_n, q_n, p_n, \theta) = \int \mathcal{P}_n(\nu) \cdot \{\Omega^{-1}(\nu - \mu)(\nu - \mu)^T \Omega^{-1} - \Omega^{-1}\} \Lambda \cdot F(d\nu | \theta)$$

A straightforward MSS approach to computation of (13)-(17) in the multivariate normal case is to start from a large repository $r = 1, \ldots, R$ of draws of i.i.d. standard normal vectors $\zeta_r$, which are held fixed during estimation to avoid chatter and maintain stochastic equicontinuity, and iteratively consider trial values $\theta = (\mu, \Lambda)$, with $\Omega = \Lambda \Lambda'$, and values $\nu_r = \mu + \Lambda \zeta_r$ for the intermediate parameter vectors $\nu$ that enter $\mathcal{P}_n(\nu)$. Then approximate (13) at $\theta$ by $\frac{\hat{L}_n}{\hat{\theta}}(d_n | x_n, q_n, p_n, \theta) = \frac{1}{R} \sum_{i=1}^{R} \mathcal{P}_n(\nu_r)$, (17) by

$$\nabla_{\mu} \hat{L}_n = \frac{1}{R} \sum_{i=1}^{R} \mathcal{P}_n(\nu_r) \Omega^{-1}(\nu_r - \mu)$$

$$\nabla_{\Lambda} \hat{L}_n = \frac{1}{R} \sum_{i=1}^{R} \mathcal{P}_n(\nu_r)\{\Omega^{-1}(\nu_r - \mu)(\nu_r - \mu)^T \Omega^{-1} - \Omega^{-1}\} \Lambda ,$$

28
and (16) by an average over n of the outer product of the gradient (18), reshaped into a vector; i.e., define $g_n$ to be a column vector that concatenates $\nabla_\mu \bar{L}_n$ and a vector containing the lower triangular elements of $\nabla_\Lambda \bar{L}_n$, column by column, and $\Sigma_N = \frac{1}{N} \sum_{n=1}^{N} g_n g_n^T$. Other treatments of the problem of estimating mixed MNL models, using MSL rather than MSS, are Ben-Akiva et al. (1993), Revelt and Train (1998), Train (1998, 1999, 2007, 2009), Brownstone and Train (1999), Brownstone, Bunch, and Train (2000), Chesher and Santos-Silva (2002), Hensher and Green (2002), Train and Sonnier (2005), and Walker et al. (2007).

Prof. Arne Rise Hole has recently developed STATA modules to estimate mixed logit models in preference space and wtp-space. The modules, called “mixlogit” for preference space and “mixlogitwtp” for wtp space, are available at https://www.sheffield.ac.uk/economics/people/hole/stata or by typing “ssc install mixlogit” and “ssc install mixlogitwtp” in the STATA command window. The two modules share syntax and include options for estimation by various maximization techniques, including Newton-Raphson and BHHH; for correlated or uncorrelated random terms (coefficients in preference space, WTP’s and alpha in wtp-space); and for normal and lognormal distributions.

When the target of the CBC is market prediction rather than estimates of individual preferences, there only a modest loss of generality or efficiency in ignoring $\alpha, \beta, \delta$ homogeneity across menus for the same consumer. Then the model (13) reduces to the mixed MNL model (4) that ignores the panel structure of stated choices from multiple menus. The simulation method described above still applies.

5. HIERARCHICAL BAYES ESTIMATION

An alternative and/or complement to classical statistical estimation is the use of Bayes methods. These turn out to be particularly convenient for analysis of CBC data, and have been refined and applied widely in marketing, led by the contributions of Allenby and Rossi (1993, 1994) and Rossi, Allenby, and McCulloch (2005, 2006). Textbook treatments of Bayesian statistics can be found in Koop (2003), and Gelman et al. (2004). Under general conditions, the mean of the Bayesian posterior of a parameter is asymptotically equivalent to the classical maximum likelihood estimator (MLE), and the variance of the Bayesian posterior is asymptotically equivalent to the sampling variance of the MLE. This equivalence implies that a researcher can implement a Bayesian estimation procedure and treat the mean and variance of the posterior the same as classical estimates and their covariances, without

---

18 Technically, the independent elements in the lower triangular array $\Lambda$ are vectorized and concatenated with $\mu$ to form the vector $\theta$; the R command for doing this is $\theta \leftarrow c(\mu, \Gamma[\text{lower.tri}(\Gamma, \text{diag=TRUE})])$. The corresponding vector $g_n$ is formed by the R command $\nabla_\Lambda \bar{L}_n[\text{lower.tri}(\nabla_\Lambda \bar{L}_n, \text{diag=TRUE})]$. 

necessarily adopting a Bayesian philosophy about statistics. This section first reviews the basics of Bayes estimation, then shows its use for CBC analysis.

5.1. Bayesian Estimation. Consider the problem of making a decision under uncertainty, when there is a cost to mistakes. For example, a firm making a pricing decision on its product is uncertain on how the price it sets will affect its product demand and profit, and the cost of a mistake is lost profit. Uncertainty comes from incomplete knowledge on consumer choice behavior. (There may also be uncertainty about strategies of rival firms, but this is omitted from the current discussion.) Suppose the firm conducts a CBC study of consumer choice. The observations $y$ from this study are realizations of a random vector that are governed by a probability law or data generation process (DGP) indexed by a parameter vector $\theta$ and described by a density or likelihood $L(y,\theta)$. For example, in a study of automobile stated purchase choices, one might specify a flat MNL data generation process, so that $L(y,\theta)$ would have the form (12). The decision task is to choose an estimator $t(y)$ of the true parameter vector $\theta$. Assume the cost of a mistake, in this case lost profit, is a function $c(t – \theta)$, so the cost of using an estimator $t(y)$ is $c(t(y) – \theta)$. You would like to select an estimator $t(y)$ that makes this expression small, but $\theta$ is unknown. A classical statistical approach is to select an estimator that is close to $\theta$ with high probability, no matter what the true value of $\theta$, without explicitly considering the cost $c(t – \theta)$. For example, the estimator $t_{\text{MLE}}(y)$ that maximizes the likelihood $L(y,t)$ for most problems will in increasingly large samples be in each small neighborhood of the true $\theta$ with a probability that approaches one. Then, with sufficient data, the expected cost of using $t_{\text{MLE}}(y)$ is small. However, the MLE has drawbacks. Its computation via numerical maximization of $L(y,t)$ can be burdensome. The fact that it is well-behaved in very large samples does not guarantee that it has particularly good properties in the finite samples encountered in applications. Finally, it is not tailored to minimize the particular cost of mistakes $c(t – \theta)$ that the decision-maker faces, nor can this be accomplished easily within the limits of classical statistics, since to evaluate $c(t – \theta)$, you face the “belling the cat” problem that you need to know $\theta$, the unknown you are trying to estimate.19

A Bayesian or decision-theoretic statistical approach instead postulates that $\theta$ can itself be treated as a random quantity. Prior to observing $y$, the decision-maker has beliefs on what values of $\theta$ are likely. These beliefs can be expressed in terms of a prior density $k(\theta)$. Then, the joint density of $y$ and $\theta$ is $L(y,\theta)k(\theta)$, and the conditional density of $\theta$ given $y$, the decision-maker’s posterior belief about $\theta$ given the data, is, by Bayes Law,

$$K(\theta|y) = \frac{L(y,\theta)k(\theta)}{\int_{\theta'} L(y,\theta')k(\theta')d\theta'}.$$

The expected cost of an estimator $t(y)$ is then

---

19 It is possible to use resampling and bootstrap methods to approximate the distribution of $c(t – \theta)$. 

30
This is called the Bayes Risk associated with the estimator \( t(y) \). The decision-maker can pick a Bayes estimator \( t(y) \) to minimize (20), and this estimator is then tailored to give the best possible solution to this decision-maker’s problem. The first-order condition for minimizing (20) is

\[
0 = \int_\theta c'(t(y) - \theta)K(\theta|y)d\theta.
\]

The Bayes estimator is a logically and intuitively appealing solution to the decision-maker’s problem. However, it has drawbacks. First, computation of the \( t(y) \) that minimizes (20) can be challenging. Second, the solution depends on the prior \( k(\theta) \) and the cost \( c(t(y) - \theta) \) of the particular decision-maker, and while they are optimal for her, they are not necessarily optimal for anyone else. Further, the decision-maker may be unsure of both her prior beliefs and her costs of mistakes. In this sense, Bayes estimators are personal and non-robust, and while it is possible to make them more generic by adopting general-purpose priors and cost functions, this also loses some of the value of these estimators to the individual decision-maker.

Bayes estimators balance the information on \( \theta \) contained in the likelihood \( L(y, \theta) \) for the observed data and that contained in the prior \( k(\theta) \). Priors can be either diffuse or tightly concentrated, and if they are tight, they can be “good”, concentrated near the true value of \( \theta \), or “bad”, concentrated far away from it. An example shows the effect of different priors. Suppose you operate a news stand, and must decide how many papers \( s \) of a particular brand to stock. You make a profit of \( r \) on each paper sold, and a loss of \( c \) on each paper that remains unsold. You believe the number of potential buyers \( k = 0, 1, \ldots \) follows a probability law \( P(k) \). Then \( \pi(s) = \sum_{k=0}^{\infty} \left( r \cdot k - c \cdot (s - k) \right) P(k) = r s - (r + c) \sum_{k=0}^{s-1} Q(s) \) is your expected profit at stock \( s \), where \( Q(s) = \sum_{k=0}^{s} P(k) \) is the CDF of this probability law. The incremental profit from increasing \( s \), \( \pi(s+1) - \pi(s) = r - (r + c)Q(s) \) is positive when \( Q(s) < r/(r+c) \). Then, the optimal stock \( s \) is the \( r/(r+c) \) quantile of \( Q \), satisfying \( Q(s - 1) \leq r/(r+c) < Q(s) \).

Suppose you are sure that the probability law is Poisson, \( P(k|\theta) = e^{-\theta} \theta^k/k! \), with a positive parameter \( \theta \). If you knew \( \theta \), then calculation of the optimal stocking level would be straightforward. For example, when \( \theta = 20 \), the Poisson density has mean 20, and its mode and median are both between 19 and 20. If \( r = 0.50 \) and \( c = 1.00 \), the optimal \( s \) is 18, the 0.33 quantile of the Poisson CDF. Alternately, if \( r = s = 1.00 \), the optimal \( s \) is 20, the median of the CDF.

In actuality, \( \theta \) is unknown. As a Bayesian decision-maker, you have prior beliefs \( k(\theta) \) on \( \theta \) and data \( y \) described in terms of its likelihood \( L(y, \theta) \). Using (19), you form the posterior \( K(\theta|y) \). You then use this result to calculate the probability \( P(k) = \int_\theta e^{-\theta} \frac{\theta^k}{k!} K(\theta|y)d\theta \), its associated CDF \( Q(k) \), and finally the stocking level \( s \).
determined by the quantile $Q(s - 1) \leq r/(r+c) < Q(s)$. The first step in these calculations is to specify your prior $\theta$ reflecting your beliefs about $\theta$. There are many possibilities, some discussed later, but it will be particularly convenient if your prior is a gamma distribution, the density $k(\theta) = \frac{\beta^\alpha \theta^{\alpha-1} e^{-\theta \beta}}{\Gamma(\alpha)}$, where $\alpha$ and $\beta$ are positive parameters. This prior has mean $\alpha/\beta$ and variance $\alpha/\beta^2$, or turning these around, $\beta = \text{mean}/\text{variance}$ and $\alpha = (\text{mean})^2/\text{variance}$. Then, keeping $\alpha/\beta$ constant, making $\beta$ large leads to a tight or concentrated prior around $\alpha/\beta$ that resembles a normal bell curve, and making $\beta$ small leads to a diffuse, uninformative prior that resembles an exponential density.

Next suppose for $N = 20$ working days, you record the number of people $y_n$ who arrive and ask for the paper; the data $y = (y_1, \ldots, y_N) = (15,25,20,19,27,22,24,23,20,16,17,13,25,18,15,17,22,12,13,19,17,19)$. Note that these are complete observations, not censored observations on the number of papers sold; with censored data one would have to calculate the density of the censored counts conditioned on the stocks available. Letting $Y \equiv \sum_{n=1}^{N} y_n = 380$ denote the sum of the observed counts, the log likelihood of the sample of complete observations is

$$\log L(y, \theta) = \sum_{n=1}^{N} \log \left[ e^{-\theta y_n / y_n!} \right] = Y \cdot \log \theta - N \cdot \theta - \sum_{n=1}^{N} \log(y_n!) .$$

This log likelihood is maximized at $t_{\text{MLE}}(y) \equiv Y/N = 19$.

Given your gamma prior and the likelihood (22), you can compute your posterior density using formula (19),

$$K(\theta|y) = \frac{L(y, \theta)k(\theta)}{\int_0^{+\infty} L(y, \theta')k(\theta')d\theta'} = \frac{(N+\beta)^{Y+\alpha} e^{-(N+\beta)\theta}}{\Gamma(Y+\alpha)}.$$ 

This is again a gamma density, so that the prior and the posterior are in the same family, and are called conjugates. Using a conjugate prior is by no means necessary for Bayesian analysis, but when available and reasonably consistent with your beliefs, it simplifies computation. From (23), you can read off the mean $(Y+\alpha)/(N+\beta)$, mode $(Y+\alpha-1)/(N+\beta)$, and variance $(Y+N + \alpha)/(N + \beta)^2$ of the posterior, and its $\gamma$ quantile is given for example by the R function qgamma(\gamma, Y+\alpha, rate=1/(N+\beta),0,1). Further, the gamma prior gives a simple form for the probability

$$P(k) \equiv \int_0^{+\infty} e^{-\theta} \frac{\theta^k}{k!} K(\theta|y)d\theta = \int_0^{+\infty} e^{-\theta} \frac{\theta^k}{k!} \frac{(N+\beta)^{Y+\alpha} e^{-(N+\beta)\theta}}{\Gamma(Y+\alpha)} d\theta = \left( k + Y + \alpha - 1 \right) \left( 1 - \frac{1}{1+N+\beta} \right)^{Y+\alpha} \left( \frac{1}{1+N+\beta} \right)^k ,$$

where $\binom{n}{k} = \Gamma(n+k+1)/\Gamma(k+1)\Gamma(n+1)$ denotes a combinatorial; (24) is a negative binomial density with mean $(Y + \alpha)/(N+\beta)$, so that the posterior expectations of $k$ and $\theta$ are the same. The CDF $Q(k)$ for the probability (24) is given
for example by the R function `dnbinom(k,size=Y+α,mean = \frac{N+β}{1+N+β})`. For the newsstand decision-maker’s problem, consider three gamma priors shown as solid curves in Figure 1, one diffuse with low information, and two concentrated at low and high levels. This figure also plots the posterior distributions corresponding to these three priors and to the likelihood (22) with \(N = 20\) and \(Y = 380\). The posteriors from the tight priors are pulled toward the maximum likelihood estimate \(t_{MLE}(y) = 19\), but to much less extent than for the diffuse prior.

**Figure 1. Prior (solid) and Posterior (dashed) Densities**

Table 3 gives the parameter values for the priors in Figure 1, and the modified parameters in the conjugate posteriors that use the sample information. The diffuse prior has a large variance, and its posterior is dominated by the likelihood, so that its posterior median is very close to the maximum likelihood estimator \(t_{MLE}(y) = 19\). For each of the three priors, the posterior means and medians are quite close. The last two rows of the table give the optimal stocking levels and expected profits when \(θ = 20\). The high tight prior, with a median near the true value, leads to the most profitable stocking decision, but the diffuse prior that leads to a posterior focused on the likelihood, is close. The low tight prior, which places high confidence around an underestimate of true demand, does substantially worse.  Conclusion: Unless you have genuinely high confidence in your beliefs, use a relatively diffuse prior.
Table 3. Alternative Gamma Conjugate Priors and Posteriors

<table>
<thead>
<tr>
<th></th>
<th>Diffuse</th>
<th></th>
<th>High Tight</th>
<th></th>
<th>Low Tight</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>0.799</td>
<td>380.8</td>
<td>170.00</td>
<td>550</td>
<td>80.35</td>
<td>460.35</td>
</tr>
<tr>
<td>( \beta )</td>
<td>0.05</td>
<td>20.05</td>
<td>7</td>
<td>27</td>
<td>8</td>
<td>28</td>
</tr>
<tr>
<td>Mean</td>
<td>15.98</td>
<td>18.98</td>
<td>24.29</td>
<td>20.37</td>
<td>10.04</td>
<td>16.44</td>
</tr>
<tr>
<td>Median</td>
<td>10.01</td>
<td>18.96</td>
<td>24.24</td>
<td>20.36</td>
<td>10.00</td>
<td>16.43</td>
</tr>
<tr>
<td>Variance</td>
<td>319.6</td>
<td>0.95</td>
<td>3.47</td>
<td>0.75</td>
<td>1.26</td>
<td>0.59</td>
</tr>
<tr>
<td>Optimal s and actual expected profit at ( r = 0.50 ) and ( c = 1.00 )</td>
<td>16</td>
<td>17</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$7.39</td>
<td>$7.56</td>
<td>$5.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimal s and actual expected profit at ( r = 1.00 ) and ( c = 1.00 )</td>
<td>18</td>
<td>19</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$16.15</td>
<td>$16.39</td>
<td>$12.84</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The newsstand example took into account the specific profit implications of alternative decisions, rather more complicated than just minimizing Bayes risk, but used computationally convenient conjugate priors. In practice, Bayesian analysis often adopts generic cost functions that are easy to work with, and uses a variety of methods to form priors and calculate posteriors. Three commonly used cost functions are illustrated in Figure 2. First, the quadratic cost function \( c(t(y) - \theta) = (t(y) - \theta)^2 \), substituted in the first-order condition (20), gives \( 0 = t(y) - \int \theta K(\theta|y)d\theta \), so \( t(y) \) is the mean of the posterior distribution. Second, \( c(t(y) - \theta) = (1 - \gamma)\max(0,t(y) - \theta) - \gamma\min(0,t(y) - \theta) \) is linear in deviations from \( \theta \), and in (17) gives \( 0 = (1 - \gamma)\int_{-\infty}^{t(y)} K(\theta|y)d\theta - \gamma \cdot \int_{t(y)}^{+\infty} K(\theta|y)d\theta \), solved when \( t(y) \) is the \( \gamma \)-quantile of the posterior distribution. If linear costs are symmetric, then \( \gamma = \frac{1}{2} \) and \( t(y) \) is the median of \( K(\theta|y) \). Third, consider the bounded cost function \( c(t(y) - \theta) = 1 - I(|t(y) - \theta| < \gamma) \). Substituted in (19), this gives the expected cost \( 1 - \int_{t(y) - \gamma}^{t(y) + \gamma} K(\theta|y)d\theta \approx 1 - 2\gamma K(t(y)|y) \), where the approximation is good when \( \gamma \) is small. This approximation is minimized when \( t(y) \) maximizes the posterior density.
Figure 2. Common Cost functions for Bayes Analysis

A feature of the newsstand example was that the MLE and posterior median and mean estimates are close, even with a moderate sample $N = 20$. This feature holds quite generally as $N$ becomes large. To see this, start from the log likelihood function

$$
\log L(y, \theta) = \sum_{n=1}^{N} \log f(y_n, \theta),
$$

where $y_n$ is the data, $f(y_n, \theta)$ is the density for observation $n$, and $N$ is sample size. Take the identity

$$
1 \equiv \int_{-\infty}^{\infty} f(y_n, \theta) \, dy_n = \int_{-\infty}^{\infty} \exp(\log f(y_n, \theta)) \, dy_n
$$

and differentiate it in $\theta$ to get

$$
0 = \int_{-\infty}^{\infty} \frac{\partial \log f(y_n, \theta)}{\partial \theta} f(y_n, \theta) \, dy_n \quad \text{and} \quad 0 = \int_{-\infty}^{\infty} \left[ \left( \frac{\partial \log f(y_n, \theta)}{\partial \theta} \right)^2 + \frac{\partial^2 \log f(y_n, \theta)}{\partial \theta^2} \right] f(y_n, \theta) \, dy_n.
$$

For regular problems, $I(\theta') = \int_{-\infty}^{\infty} \left( \frac{\partial \log f(y_n, \theta')}{\partial \theta} \right)^2 f(y_n, \theta) \, dy_n$, termed the Fisher Information, is positive for $\theta'$ near $\theta$. A Taylor’s expansion around the true value $\theta_0$ of $\theta$ gives

$$
K(\theta|y)/K(\theta_0|y) = \exp[S_N \cdot \sqrt{N}(\theta - \theta_0) - \frac{1}{2} I_N \left( \sqrt{N}(\theta - \theta_0) \right)^2 + \frac{1}{\sqrt{N}} k' \cdot \sqrt{N}(\theta - \theta_0)].
$$

By a central limit theorem, $S_N = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \frac{\partial \log f(y_n, \theta_0)}{\partial \theta}$ converges to a normal random variable with mean zero and variance $I(\theta_0)$. By a law of large numbers and additional technical arguments, $I_N = \frac{1}{N} \sum_{n=1}^{N} \left( \frac{\partial \log f(y_n, \theta)}{\partial \theta} \right)^2$ converges to $I(\theta_0)$. Then, for $N$ large, the argument in the exponent on the right-hand-side of (28) behaves like $\sqrt{N}(\theta - \theta_0)$ times the expression $S_N + \frac{1}{\sqrt{N}} k' - \frac{1}{2} I_N \sqrt{N}(\theta - \theta_0)$. But this expression has the negative of the sign of $\sqrt{N}(\theta - \theta_0)$

35
when \( \sqrt{N}(\theta - \theta_0) \) is large in magnitude, establishing from (28) that the posterior mode has \( \sqrt{N}(\theta - \theta_0) \) bounded. Further, the contribution of the prior, \( \frac{1}{\sqrt{N}} k' \), is asymptotically negligible, so that the mode of the posterior and the MLE converge and have the same asymptotic distribution. Finally, the concentration of the density \( K(\theta|\mathbf{y}) \) in a neighborhood where \( \sqrt{N}(\theta - \theta_0) \) is bounded, along with an assumption that the tails of \( k(\theta) \) are sufficiently thin so that the mean of the posterior is not dominated by its tails, imply that the mean of \( K(\theta|\mathbf{y}) \) and its \( \gamma \)-quantiles for \( \gamma \) in the interior of \((0,1)\) are all in a \( \sqrt{N}(\theta - \theta_0) \) bounded neighborhood of \( \theta_0 \), and hence \( \frac{1}{\sqrt{N}} \) close to \( \theta_0 \) and the MLE for \( N \) large. More precise statements can be found in Ibragimov and Has'minskii (1973) and Strasser (1975). Huber and Train (2001) find that in a dataset of typical size, the MLE and Bayes procedures produce very similar estimates for a mixed logit model in preference space. Scarpa et al. (2008) obtains similar results from Bayesian and MLE for mixed logit models in willingness-to-pay space. Train (2009) provides further discussion and examples.

5.2. Selecting Priors and Computing Posteriors. The newsstand example above with Poisson data used a gamma conjugate prior, substantially simplifying calculation of the (Gamma) posterior and statistics from it such as the posterior mean. Statistical problems with natural conjugate priors are special, but convenient when applicable. Fink (1997) describes the conditions under which problems admit conjugate priors, notably distributions in the exponential family, and discusses various transformations of problems that allow conjugate priors to be used. Gelman et al. (2003) gives a compendium of known priors; see also the Wikipedia entry on conjugate priors. Of value in CBC applications are the Dirichlet prior that is conjugate for a multinomial likelihood\(^{20}\), and the Normal/Wishart prior that is conjugate for a multivariate normal likelihood with unknown mean and covariance matrix\(^{21}\). Drawbacks of concentrating on conjugate priors are first that even if the posterior density is analytically determined, it is not necessarily simple to compute moments or draw simulated values from it, and second that natural parameterizations of models for CBC data often do not admit conjugate priors.

Usually a CBC analysis can start with a prior density and a choice model likelihood that can be expressed in terms of computationally tractable functions. Even so, except for special cases of conjugate priors, it is usually impossible to obtain analytically the mean \( \int_{\theta} \theta \cdot K(\theta|\mathbf{y})d\theta \) or other key characteristics of the posterior density, or

\(^{20}\) If the likelihood of counts \((k_1,\ldots,k_J)\) from \( N \) trials is \( \binom{N}{k_1 \ldots k_J} p_1^{k_1} \cdots p_J^{k_J} \) and the prior is the Dirichlet density \( \frac{\Gamma(\alpha_1+\cdots+\alpha_J)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_J)} p_1^{\alpha_1-1} \cdots p_J^{\alpha_J-1} \), then the posterior is Dirichlet with parameters \( \alpha_j + k_j \).

\(^{21}\) If the likelihood is formed from \( n \) draws from a normal \( \mathcal{N}(\mu,\Sigma^{-1}) \) density with unknown \( \mu \) and \( \Sigma \), the prior for \( \mu \) is normal, and the prior for \( \Sigma^{-1} \) is independently inverse Wishart, then the posterior distribution for \( \mu \) is multivariate normal and for \( \Sigma^{-1} \) is independently inverse Wishart. There are alternative parameterizations and conjugate priors for \( \Sigma \) that may have better computational and statistical properties. For example, an LU decomposition expresses \( \Sigma = LL' \), where \( L \) is lower triangular with a positive diagonal, with a normal/variance-gamma conjugate prior.
even to compute easily the scaling factor $A = \int \theta \cdot L(\theta | y)k(\theta)d\theta$ that makes $K(\theta | y) = L(\theta | y)k(\theta)/A$ integrate to one. However, with contemporary computers it is practical in most CBC applications to use some combination of numerical integration and simulation methods to estimate features of the posterior density such as its mean. There is a large literature on methods for these calculations and their computational efficiency; see for example Abramowitz and Stegum (1964, chap 25.4), Press et al. (1986), Quarteroni et al. (2007, Chap 9), Brooks et al. (2011, Chap 5), Hoffman and Gelman (2014). This paper will describe several basic methods, particularly those that have proven useful for Bayes estimation, but we refer back to these sources readers interested in mathematical and computational details.

Deterministic numerical integration is often practical when the dimension of $\theta$ is low. Consider integrals

\begin{equation}
A \cdot E(\theta | y) \equiv \int_0^{\Theta} \theta^j \cdot L(\theta | y)k(\theta)d\theta = \int_0^{\Theta} \theta^j \cdot \left\{ \frac{L(\theta | y)k(\theta)}{h(\theta)} \right\} h(\theta)d\theta,
\end{equation}

for $j = 0, 1$, with the last formula obtained by multiplying and dividing by a positive function $h(\theta)$ selected to facilitate computation. For example, $h$ may be a probability density with a CDF that is easy to compute. Quadrature is a standard approach to computing (29): For a finite collection of points $\theta_i$, $i = 1, \ldots, I$, and associated weights $h_i$, approximate (29) by $\sum_{i=1}^{I} \theta_i^j \cdot \left\{ \frac{L(\theta | y_i)k(\theta_i)}{h(\theta_i)} \right\} h_i$. For example, the $\theta_i$ might enumerate the centroids of a finite grid that partitions $\theta$-space, with $h_i$ the probability that a random vector with density $h$ falls in the partition with centroid $\theta_i$. This is the computational analogue of Reimann integration, with accuracy limited by the degree of variation of the integrand $L(\theta | y)k(\theta)/h(\theta)$ on the partition; e.g., the accuracy with which the integrand can be approximated by step functions. Computation efficiency is improved by selecting a practical $h$ such that the integrand varies slowly with $\theta$. There are a variety of mathematical methods that exploit smoothness features of the integrand to improve computationally efficiency; e.g., Quarteroni et al. (2007,9.1-9.7). A general idea behind quadrature is that various series of functions other than step functions, such as orthogonal polynomials or splines, have analytic integrals with respect to $h$, and points and weights can be selected so that approximation above to (26) is exact for the initial terms in the specified series. For example, a widely used method termed Gaussian quadrature selects points and weights such that the approximation above is exact for the family of Hermite orthogonal polynomials up to a given order. This can be computationally quite efficient, but two drawbacks are that the number of functional evaluations required rises to the power of the dimension of $\theta$ and quickly becomes intractable, and that the approximation is poor when the integrand varies rapidly relative to the approximating Hermite polynomials.

A simple simulation method for approximating (29) is importance sampling. Start from a density $h(\theta)$ from which it is computationally easy to draw random deviates and for which $L(\theta | y)k(\theta)/h(\theta)$ is bounded for all $\theta$. An average of the integrand in (29) over a sample of simulated independent random deviates from $h$ will by a law of
large numbers converge to the integral.\textsuperscript{22} The primary drawback of importance sampling is that unless the bracketed integrand is quite smooth and slowly varying in $\theta$, the estimates will be relatively inefficient (as measured by precision achieved per unit of computational resources). However, variance reduction methods such as quota sampling from partitions of $\theta$-space and use of antithetic variates that exploit symmetries in $h$ can improve efficiency; e.g., Rubinstein et al. (1985). There are a variety of simulation methods that may improve substantially on the computational efficiency of importance sampling, some utilizing mathematical analogies to conservation of energy in physics, tempering and annealing in metallurgy, and minimization of entropy in probability theory. We explain several simple methods that are widely used in Bayesian analysis, and conclude with a survey of more advanced methods.

One simple alternative to the i.i.d. sampling from the density $h$ used in importance sampling is termed acceptance/rejection (A/R) sampling. This method gives a sample from the posterior $K(\theta|y)$ without requiring estimation of the normalizing scale factor $A = \int \theta L(y|\theta)k(\theta)d\theta$. Let $f(\theta|y) = L(y|\theta)k(\theta) \propto K(\theta|y)$, and find a comparison density $h(\theta)$ from which it is easy to make pseudo-random draws, with the property $Bh(\theta) \geq f(\theta|y)$ for some positive scalar $B$.\textsuperscript{23} Have the computer draw a sequence of pairs of pseudo-random numbers $(\theta_i, u_i)$, where $\theta_i$ is a draw from $h$ and $u_i$ is a draw from a uniform density on $[0,1]$. Accept and retain $\theta_i$ if $u_iBh(\theta_i) \leq f(\theta_i|y)$; otherwise, reject and discard it. Then

$$\text{Prob}(\theta|\text{accept}) = \frac{\text{Prob}(\theta \& \text{accept})}{\int \text{Prob}(\theta' \& \text{accept})d\theta'} = \frac{h(\theta) \cdot f(\theta|y)/Bh(\theta)}{\int h(\theta') \cdot f(\theta'|y)/Bh(\theta')d\theta'} = \frac{f(\theta|y)/B}{\int f(\theta'|y)/Bd\theta'} = K(\theta|y),$$

and the sequence of accepted points $\theta_i$ will be a random sample from $K(\theta|y)$ as claimed. To illustrate, consider a choice experiment in which $n = 1, \ldots, N$ subjects are asked if they will purchase a product at price $x_n$, and $d_n$ is an indicator that is one if the response is “Yes”, zero otherwise; then $y_n \equiv (x_n, d_n)$. Suppose the probability model is logistic, $\text{Prob}(y_n|x_n) = \exp(-d_n x_n \theta)/(1+\exp(-x_n \theta))$, and the prior is standard exponential, $e^\theta$ for $\theta < 0$. Then $K(\theta|y) \propto \ldots$

\textsuperscript{22} When $\theta_j \cdot \frac{L(y|\theta_j)k(\theta)}{h(\theta)}$ is bounded by a constant $C$ for $j = 0,1$, Hoeffding’s inequality (Pollard, 1984, Appendix B) establishes that $\text{Prob}\left(\sum_{i=1}^n \theta_i \cdot \frac{L(y|\theta_i)k(\theta_i)}{h(\theta_i)} - A \cdot \mathbf{E}(\theta|y) \right) > \eta < 2 \cdot e^{-2n^2/2C^2}$. Then with some manipulation, one can show that for $\eta < A/2$, $\text{Prob}\left(\sum_{i=1}^n \theta_i \cdot \frac{L(y|\theta_i)k(\theta_i)}{h(\theta_i)} - \mathbf{E}(\theta|y) \right) > \eta < 4 \cdot e^{-\eta^2/32\cdot\max(C^4,1)}$, so convergence is exponential.

\textsuperscript{23} In the R core statistical package, vectors of draws from the beta, binomial, Cauchy, Chi-squared, Exponential, F, Gamma, Hypergeometric, Logistic, Log-normal, Multinomial, negative binomial, normal, Poisson, T, uniform, and Weibull distributions are obtained from the respective functions rbeta, rbinom, rcauchy, rchisq, rexp, rf, rgamma, rhyper, rlogis, rnorm, rmultinom, rbinom, rmorm, rpois, rt, runif, rweibull, with arguments specifying distribution parameters. These functions can also be used, with prefixes “d”, “p”, or “q” substituted for “r”, to get densities, CDF’s, or quantiles, respectively. Additional distributions are available in packages from CRAN.
\[ f(\theta|y) = \prod_{n=1}^{N} \frac{\exp(-y_n x_n \theta)}{1 + \exp(-x_n \theta)} \cdot e^\theta 1(\theta < 0), \] an expression whose integral does not have a closed form. Note that \( f(\theta|y) \leq e^\theta 1(\theta < 0) \) and take \( h(\theta) = e^\theta 1(\theta < 0) \) and \( B = 1 \).

Figure 3 illustrates how the acceptance/rejection method works in this case. In this figure, points are drawn with equal probability from the area below the blue curve that plots \( Bh(\theta) \), extending indefinitely to the left, and are accepted if they lie below the red curve that plots \( f(\theta|y) \), with the scale factor \( B \) set so that the red curve is never higher than the blue curve. The efficiency or yield of the procedure, the share of sampled points accepted, is given by the ratio of the area under the red curve to the area under the blue curve.

![Figure 3. Illustration of Acceptance/Rejection Method](image)

A limitation of the simple acceptance/rejection method is that it can have low efficiency unless the \( h \) curve and scale factor \( B \) can be tuned to \( f(\theta|y) \); this is particularly true when the parameter vector is of high dimension. There are several modifications to A/R methods that may attain higher yields. First, in acceptance/rejection sampling, it may be possible to partition the support of \( K(\theta|y) \) into a finite number of intervals or rectangles \( A_j \), and define a value \( B_j \) that is a fairly tight upper bound on \( f(\theta|y)/h(\theta) \) for \( \theta \in A_j \). Define weights \( w_j = \frac{h(A_j)/B_j}{\sum_{j'} h(A_{j'}/B_{j'})} \). Draw \( \theta \) from \( h \) and \( u \) from uniform \([0,1]\), determine the rectangle \( A_j \) in which \( \theta \) lies, and accept \( \theta \) with weight \( 1/w_j \) if \( uB_j f(\theta|y) \leq h(\theta) \). Then the weighted accepted sample has an empirical weighted CDF that converges to the CDF of \( K(\theta|y) \), and the yield is high when the bounds within each rectangle are good. Second, instead of considering a single uniform \( u \), consider a randomly located grid composed of all values of the form \( u' = u \pm 1/r \) that fall in \([0,1]\), where \( r \) is a specified integer. Accept a value \( \theta \) with weight \( 1/w_j \) each time \( u'B_j f(\theta|y) \leq h(\theta) \) for one of these \( u' \). This modification will increase the yield as \( r \) increases, at the cost of introducing serial correlation of accepted \( \theta \)'s within blocks of length up to \( r \). Third, the \( B_j \) can be adjusted adaptively. If a “burn-in” sample of \( \theta \)'s falls in rectangle \( j \), then the maximum of \( f(\theta|y)/h(\theta) \) in this sample is a consistent estimator of the best, or least upper bound, \( B_j \). Using this best estimate may under-accept \( \theta \) from regions where the accepted subsample maximum is less than
the true least upper bound, but adaptively adjusted \(\beta_j\) will convergence very fast (e.g., at a \(1/N\) rather than a \(1/\sqrt{N}\) rate), so the bias will be negligible in simulations of reasonable size. Fourth, the partition into rectangles \(A_j\) can be refined recursively and adaptively to reduce the gap between the maximum and the minimum of \(f(\theta|y)/h(\theta)\) in each rectangle. If these refinements are made as a rate that is asymptotically negligible relative to the accepted sample size, they will introduce only asymptotically negligible noise into the simulation.

A variant of A/R sampling is termed slice sampling. It introduces an auxiliary variable \(\tau\) with a uniform density on \([0,L(y|\theta)k(\theta)]\) and a density \(\pi(\theta|\tau) = 1(\tau \leq L(y|\theta)k(\theta))\). Starting from \(\theta_{-1}\), the sampler first draws \(\tau\) from \([0,L(y|\theta_{-1}k(\theta_{-1}))\], next places a random rectangle in \(\theta\)-space around the point \((\tau,\theta_{-1})\), steps out this rectangle until it covers the set \(\{\theta | \tau \leq L(y|\theta)k(\theta)\}\), and then samples \(\theta\) from the resulting rectangle until \(\tau \leq L(y|\theta)k(\theta)\) is satisfied. In Figure 3, this corresponds to drawing horizontal strips at random, and then sampling from the portion of these strips that lie under the curve. An interpretation of slice sampling is that it is a stochastic version of Lebesgue integration, while simple A/R sampling is a stochastic version of Reimann integration.

Additional efficiency may be achievable using Monte Carlo Markov Chain (MCMC) methods; see Albert and Chib (1996), Chib and Greenberg (1996), McFadden (1996). A Metropolis-Hastings sampler is a Markov process that recursively draws points from the parameter space \(\Theta\) using a positive conditional density \(q(\theta|\theta_{-1})\) from which it is convenient to sample. Given \(\theta_{-1}\), draw a trial \(\theta^*\) from \(q(\theta^*|\theta_{-1})\) and define \(\gamma(\theta^*, \theta_{-1}|y) = \min\left(1, \frac{f(\theta^*|y)q(\theta_{-1}|\theta^*)}{f(\theta_{-1}|y)q(\theta^*|\theta_{-1})}\right)\). A key feature of the Metropolis-Hastings sampler is the symmetry of the function \(\gamma(\theta^*, \theta_{-1}|y)\) that enters the acceptance criterion. This makes the sampler reversible in time, a defining characteristic of a successful sampler. This critical value reduces to \(\gamma(\theta^*, \theta_{-1}|y) = \min(1, f(\theta^*|y)/f(\theta_{-1}|y))\) when \(q(\theta|\theta_{-1})\) is symmetric, the original, more restrictive Metropolis sampler; e.g., \(\theta - \theta_{-1}\) multivariate normal with mean zero. If a random scalar \(u\) from a uniform density on \([0,1]\) satisfies \(u < \gamma(\theta^*, \theta_{-1}|y)\), set \(\theta = \theta^*\); otherwise set \(\theta = \theta_{-1}\). This defines a Markov transition probability

\[
P(\theta|\theta_{-1}) = \begin{cases} 
q(\theta|\theta_{-1})\gamma(\theta, \theta_{-1}|y) & \text{if } \theta \neq \theta_{-1} \\
1 - R(\theta_{-1}) & \text{if } \theta = \theta_{-1} 
\end{cases}
\]

where \(R(\theta_{-1}) = \int_{\theta^0 \in \Theta} q(\theta^0|\theta_{-1}) \cdot \gamma(\theta^0, \theta_{-1}|y) \, d\theta^0\) is the probability of acceptance. The following argument shows that the posterior \(K(\theta|y)\), which equals \(A \cdot f(\theta|y)\) for some constant \(A\), is an invariant of this Markov process. If \(\theta_{-1}\) has density \(A \cdot f(\theta_{-1}|y)\), the joint density of \(\theta\) and \(\theta_{-1}\) is \(P(\theta|\theta_{-1}) \cdot A \cdot f(\theta_{-1}|y)\). Then, the marginal density of \(\theta\) is
\( A \cdot \int_{\theta_{-1} \in \Theta} f(\theta_{-1} | y) q(\theta_{-1} | \theta) \cdot \gamma(\theta, \theta_{-1} | y) d\theta_{-1} + \mathbf{1}(\theta = \theta_{-1}) \cdot A \cdot f(\theta_{-1} | y) \cdot (1 - R(\theta_{-1})) \)

\[
= A \cdot \int_{\theta_{-1} \in \Theta} f(\theta | y) q(\theta_{-1} | \theta) \cdot \min(1, \frac{f(\theta_{-1} | y) q(\theta_{-1} | \theta)}{f(\theta | y) q(\theta | \theta_{-1})}) d\theta_{-1} + \int_{\theta_{-1} \in \Theta} \mathbf{1}(\theta = \theta_{-1}) \cdot A \cdot f(\theta_{-1} | y) \cdot (1 - R(\theta_{-1}))
\]

\[
= A \cdot f(\theta | y) - R(\theta) + A \cdot f(\theta | y) \cdot (1 - R(\theta_{-1})) = A \cdot f(\theta | y),
\]

and \( K(\theta | y) \) is an invariant density. In (32), \( \mathbf{1}(\theta = \theta_{-1}) \) denotes a unit probability on the set of points in \( \Theta \times \Theta \) that meet the condition. The requirement that \( q(\theta | \theta_{-1}) > 0 \) on \( \Theta \times \Theta \), plus technical conditions on the tails of \( q(\theta | \theta_{-1}) \) relative to \( f(\theta | y) \) when \( \Theta \) is not compact, are sufficient to assure that there is a unique invariant density and the Markov process converges to this invariant density.\(^{24} \)

A drawback of MCMC methods is that they can have high serial correlation, so that very large numbers of draws and substantial “burn-in” iterations are needed to get reliable empirical approximations to the posterior.

It is fairly common in applications to require a sample from a computationally intractable multivariate distribution whose one-dimensional conditional distributions are relatively manageable. A specialization of the Metropolis-Hastings sampler called \textit{Gibbs sampling} can be used in this problem. Suppose \( \theta \) is of dimension \( d \), and \( \theta = (\theta_1, \ldots, \theta_d) \) has a multivariate density \( g(\theta_1, \ldots, \theta_d) \) with univariate conditionals \( g_i(\theta_i | \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d) \) from which it is computationally tractable to sample using, say, the quantile transformations of these densities, acceptance/rejection sampling, or a MCMC procedure. From an initial vector \( \theta' = (\theta'_1, \ldots, \theta'_d) \), the sampler loops through draws of \( \theta_i' \) from \( g_i(\theta_i | \theta_1', \ldots, \theta_{i-1}', \theta_{i+1}', \ldots, \theta_d') \) for \( i = 1, \ldots, d \). The procedure then proceeds recursively, starting next from the vector \( \theta'' \). The sampler can be started from any initial \( \theta \) vector. A leading example is a truncated multivariate normal distribution \( g \) whose one-dimensional conditionals \( g_i \) are (truncated) univariate normal. To illustrate, in the R statistical package draws from \( g_i \) subject to a truncation \( \theta_i \in [c_i, +\infty) \) are given by \( \text{qnorm}(\text{runif}(1, \text{pnorm}(c_i, \mu_i, \sigma_i, 1, 0), 1, \mu_i, \sigma_i, 1, 0)) \).

There has been considerable study of the conditions under which Metropolis-Hastings and Gibbs sampling methods perform poorly; see Wu and Fitzgerald (1996), Neal (2011), Hoffman et al. (2014), Betancourt and Girolami (2013). When \( K(\theta | y) \) has a strong gradient relative to the conditional density \( q(\theta | \theta_{-1}) \) in the Metropolis-Hastings sampler, or large differences between conditional and marginal variances in Gibbs sampling, sampler output can resemble local random walks that are highly serially correlated and fail to span the \( \theta \)-space well in a reasonable number of iterations. The problem is particularly critical when \( K(\theta | y) \) has several isolated sharp peaks.

\(^{24} \) The conditions that must be met on \( q(\theta | \theta_{-1}) \) and \( \Theta \) are \textit{irreducibility} (i.e., the process can always go from one point in \( \Theta \) to any other point in a finite number of steps), \textit{acyclicity} (i.e., the process does not always return to its starting point in a finite number of steps), and \textit{positive recurrent} (i.e., the expected time until a return to a subset of positive measure is finite); see Tirney (1994), McFadden (1996).
One straightforward method of dealing with this problem is called **annealing**, a term borrowed from metallurgy where a material is repeatedly heated above its crystallization temperature to allow atoms to migrate, and then slowly cooled. The analogy here replaces the target \( f(\theta|y) = L(y|\theta)k(\theta) \) by \( f_\tau(\theta|y) = f(\theta|y)^\tau \), which is flatter than \( f(\theta|y) \) for \( \tau < 1 \), and then over the course of the iterations periodically allows \( \tau \) to drop below one for a few iterations. Used in a Metropolis-Hastings sampler with \( \theta \)'s selected from parts of the chain where \( \tau = 1 \), this will lead a broader span of accepted values. However, the frequency, magnitude, and duration of \( \tau < 1 \) deviations, and “burn-in” times following these deviations, all have to be selected to tune the sampler to the application.

The most sophisticated versions of Monte Carlo Markov Chain methods currently in wide use are Hamiltonian Monte Carlo methods, particularly the No-U-Turn Sampler (NUTS) of Hoffman and Gelman (2014). These methods accomplish the same things as well-tuned annealing, and utilize a combination of mathematical theory and tinkering to automate some or all of the tuning process. A good introduction to these methods is Neal (2011). Hamilton Monte Carlo methods introduce auxiliary variables, in this case a vector \( \omega \) that is of the same dimension as \( \theta \), two functions \( V(\theta) = -\log(L(y|\theta)k(\theta)) \) and \( W(\omega) = \omega'M^{-1}\omega/2 \), where \( M \) is a positive definite matrix, and dynamic equations of motion in continuous time,

\[
\frac{d\theta}{dt} = \partial W(\omega)/\partial \omega = M^{-1}\omega \quad \text{and} \quad \frac{d\omega}{dt} = -\partial V(\theta)/\partial \theta = \partial \log(L(y|\theta))/\partial \theta + \partial \log(k(\theta))/\partial \theta.
\]

Interpret \( \omega \) as a momentum vector, \( W(\omega) \) as kinetic energy, and \( V(\theta) \) as potential energy. This dynamic system is time reversible and preserves the Hamiltonian \( H(\theta,\omega) = V(\theta) + W(\omega) \); i.e. \( dH(\theta,\omega)/dt = 0 \). This is analogous to conservation of energy in physical systems. The dynamic system moves rapidly out of regions where potential energy (i.e., posterior probability) is low, and slowly out of regions where potential energy is high. It has a strong mathematical property, called **symplecticness**, that the Jacobian of the transformation from \( (\theta,\omega) \) at time \( t \) to time \( t+s \) is of magnitude one; see Neal (2011, 116). An implication of this property is that this dynamic system is probability preserving; i.e., it will move along contours of constant probability density. In computational approximation to this dynamic system, continuous time is replaced by discrete time moving in small steps \( \Delta \). Start from \( \theta(t) = \theta_1 \) and a new independent draw \( \omega(t) = M^{1/2}\xi \), where \( \xi \) is a draw of a standard normal vector, independent of the previous state. Advance \( \theta \) and \( \omega \) for \( s = 0,\ldots,S-1 \) steps using what is termed the **leapfrog method**:

\[
\omega(t+(s+\frac{1}{2})\Delta) = \omega(t+s\Delta) - (\Delta/2)\cdot\partial V(\theta(t+s\Delta))/\partial \theta,
\]
\[
\theta(t+(s+1)\Delta) = \theta(t+s\Delta) + \Delta\cdot M\omega(t+(s+\frac{1}{2})\Delta),
\]
\[
\omega(t+(s+1)\Delta) = \omega(t+(s+\frac{1}{2})\Delta) - (\Delta/2)\cdot\partial V(\theta(t+(s+1)\Delta))/\partial \theta.
\]
The half-steps have to be computed only for steps 1 and S, since in between the final half-step for s coincides with the initial half-step for s+1. With this method, the discrete time approximation retains the time-reversibility and symplecticness properties of the continuous time process. At the end of the S steps, one has a proposed state \((\theta(t+S\Delta),\omega(t+S\Delta))\). This proposal is accepted as the next state if a uniform \(u\) drawn from \([0,1]\) satisfies

\[
(35) \quad u < \min\{1, \exp(V(\theta(t)) + W(\omega(t)) - V(\theta(t+S\Delta)) - W(\omega(t+S\Delta)))\};
\]

otherwise the process starts again from the previous value \(\theta_{-1}\).

The primary limitation of this Hamiltonian Monte Carlo method is that the step size \(\Delta\) and path length \(S\) have to be specified through tuning of the sampler to the application; Neal (2011) gives an extensive discussion of methods and properties in his section 5.4, and of alternatives to and variations on Hamiltonian Monte Carlo in his section 5.5. The NUTS procedure of Hoffman and Gelman (2014) is an extension of Hamiltonian Monte Carlo that makes the specification of \(S\) automatic and adaptive. The key idea is to stop the leapfrog iteration (34) when the Euclidean distance between \(\theta(t)\) and \(\theta(t+s\Delta)\) stops increasing at step \(s\), a “U-turn”. To preserve time reversibility, NUTS uses an algorithm that runs forward or backward one step, then forward or backward two steps, and so on until a U-turn occurs. At that point, it samples from the points collected through the iteration in a way that preserves symmetry, and applies a criterion like (35) to determine if the selected point is accepted. Hoffman and Gelman (2014) give a detailed discussion of this method, including proof of its time reversibility properties, recommendations and code for its implementation, a method for adaptively setting the step size \(\Delta\), and empirical experience with NUTS for Bayesian analysis of some standard test data. On the basis of the reported performance of this algorithm, and its availability in the open-source, R-compatible STAN statistical package developed by the authors and associates (Carpenter et al., 2015), we believe it to be the best method currently available for Bayesian analysis, particularly hierarchical Bayes analysis of CBC studies.

5.3. Hierarchical Bayes Methods. In marketing, a widely used model for CBC data is a mixed (or random coefficients) multinomial logit model like (4) for the choice probabilities, with a Bayesian prior \(k(\theta)\) on the deep parameters; see Rossi et al. (2005), Ter Hofstede et al. (2002), Lenk et al. (1996). Let \(L_n(d_n|x_n, q_n, p_n, \nu)\) denote the likelihood, from (13), that respondent \(n\) will state a vector of choices \(d_n\) from the offered menus. Then the posterior density on these parameters is

\[
(36) \quad K(\theta|x, p) \propto \prod_{n=1}^{N} \int_{\nu} L_n(d_n|x_n, q_n, p_n, \nu) \cdot F(d\nu|\theta) \cdot k(\theta).
\]

\(25\) The symmetry of the Hamiltonian process requires that the sign of \(\omega(t+S\Delta)\) be reversed if the proposed state is accepted. However, this step is not needed unless this process is interwoven with other Monte Carlo methods.
To illustrate the use of (36) to produce Bayesian estimates of its parameters, consider a specification with a fairly conventional relatively uninformative prior \( k \) and a density \( f \) that is multivariate normal except for a lognormal transformation or censoring in some components, as described by Sonnier and Train (2005) and summarized in the discussion following (13). Let \((\alpha, \beta, \delta)\) be a transformation of a commensurately-sized latent vector \( \nu \) of normally distributed terms with mean \( \mu \) and covariance matrix \( \Omega \), so that the parameters to be estimated are \( \theta = (\mu, \Omega) \). Alternately, write \( \Omega = \Lambda \Lambda' \), where \( \Lambda \) is a lower triangular matrix. In this case, \( \theta = (\mu, \Lambda) \). In either case, if \( t \) is the dimension of \((\alpha, \beta, \delta)\), then the dimension of \( \theta \) is \( t(t+3)/2 \); this is also the dimension of \( \nu \). The procedure will repeatedly use the following operation for drawing \((\alpha, \beta, \delta)\), referred to hereafter as the *taste parameter draw*: Given a trial value \( \theta \) and a vector \( \zeta \) of iid standard normal deviates, let \( \nu = \mu + \Omega^{1/2} \zeta \equiv \mu + \Lambda \zeta \) be a \( t \times 1 \) multivariate normal vector. Each component of \((\alpha, \beta, \delta)\) that is unrestricted in sign is set equal to the corresponding component of \( \nu \); components like \( \alpha \) that are necessarily positive are set to the exponent of the correspondent element of \( \nu \); and components that are censored to be non-negative are set to the maximum of 0 and the correspondent element of \( \nu \). Other transformations are of course possible. We describe two widely used procedures that have been applied for this specification.

*Allenby-Train Procedure*: Assume that the prior \( k(\theta) \) is the product of (a) multivariate normal density for \( \mu \) with zero mean and sufficiently large variance so that the density is flat from a numerical perspective, and (b) an inverted Wishart for \( \Omega \) with \( T \) degrees of freedom and parameter \( TI_T \), where \( I_T \) is a \( T \)-dimensional identity matrix.

Draws from the posterior distribution are obtained by iterative Gibbs sampling from three conditional posteriors, with a Monte Carlo Markov Chain algorithm used for one of the conditional posteriors. Let \( \mu_i, \Omega_i, \) and \( \nu_{ni} \) be values of the parameters in iteration \( i \), with the taste parameter draw implied by \( \nu_{ni} \). Recall that \( N \) is the number of subjects. Gibbs sampling from the conditional posteriors for each parameter is implemented in three layers:

\[
\begin{align*}
\mu \mid \Omega, \nu_n & \text{ for all } n: \text{Given } \Omega_i \text{ and } \nu_{ni} \text{ for all } n, \text{ the conditional posterior on } \mu \text{ is } N\left(\bar{\nu}_i, \frac{\Omega_i}{N}\right) \text{ where } \bar{\nu}_i = \frac{1}{N} \sum_n \nu_{ni}. \text{ A new draw of } \mu \text{ is obtained as } \mu_{i+1} = \bar{\nu}_i + \Psi_i \eta \text{ where } \Psi_i \text{ is the lower-triangular Cholesky factor of } \Omega_i/N \text{ and } \eta \text{ is a draw of standard normal deviates.} \\
\Omega \mid \mu, \nu_n & \text{ for all } n: \text{Given } \mu_{i+1} \text{ and } \nu_{ni} \text{ for all } n, \text{ the conditional posterior on } \Omega \text{ is Inverted Wishart with degrees of freedom } T+N \text{ and parameter } TI+N\bar{\nu}, \text{ where } \bar{\nu} = \frac{1}{N}(\nu_i - \mu_i+1)(\nu_i - \mu_i+1)' \text{. Take } T+N \text{ draws of } T\text{-dimensional vectors of iid standard normal deviates, labeled } \eta_r, r=1,\ldots,T+N. \text{ A new draw of } \Omega \text{ is obtained as } \Omega_{i+1} = (\sum_r (\Gamma_{\eta_r})(\Gamma_{\eta_r})')^{-1} \text{ where } \Gamma \text{ is the Cholesky factor of } (TI + N\bar{\nu})^{-1}. 
\end{align*}
\]
Given \( \mu_{i+1} \) and \( \Omega_{i+1} \), for taste parameters \( \alpha_n, \beta_n, \delta_n \) implied by \( \nu_n \), the conditional posterior of \( \nu_{n,i+1} \) is proportional to \( L_n \left( \mathbf{d}_n | \mathbf{x}_n, \mathbf{q}_n, \mathbf{p}_n, \nu_{n,i+1} \right) \phi \left( \nu_{n,i+1} | \mu_{i+1}, \Omega_{i+1} \right) \), where \( \phi \) denotes the multivariate normal density with mean \( \mu_{i+1} \) and covariance matrix \( \Omega_{i+1} \), and \( L_n \) is given in (13). A draw of \( \nu_{n,i+1} \) is obtained by the following Metropolis-Hastings procedure: Draw \( s_n \) from a jumping distribution, which we specify as normal with mean zero and variance \( m \Omega \), where scalar \( m \) is chosen as described below. Calculate a trial value of \( \nu_n \) as \( \bar{\nu}_n = \nu_{n_i} + s_n \). Using the taste parameters \( (\bar{\alpha}_n, \bar{\beta}_n, \bar{\delta}_n) \) implied by \( \bar{\nu}_n \), calculate the probability \( L_n \left( \mathbf{d}_n | \mathbf{x}_n, \mathbf{q}_n, \mathbf{p}_n, \bar{\nu}_n \right) \). If a \([0,1]\) uniform random variate \( u \) satisfies

\[
    u < \frac{L_n \left( \mathbf{d}_n | \mathbf{x}_n, \mathbf{q}_n, \mathbf{p}_n, \nu_{n,i+1} \right) \phi \left( \nu_{n,i+1} | \mu_{i+1}, \Omega_{i+1} \right)}{L_n \left( \mathbf{d}_n | \mathbf{x}_n, \mathbf{q}_n, \mathbf{p}_n, \nu_{n,i} \right) \phi \left( \nu_{n_i} | \mu_{i+1}, \Omega_{i+1} \right)},
\]

then the new value of \( \nu_{n,i+1} \) is this trial \( \bar{\nu}_n \); otherwise, discard the trial \( \bar{\nu}_n \) and retain the old value as the new value, \( \nu_{n,i+1} = \nu_{n_i} \). The value of \( m \) is adjusted over iterations to maintain an acceptance rate near 0.30 over \( n \) and iterations.\(^{26}\)

**NUTS procedure:** This procedure uses the NUTS sampler developed by Hoffman and Gelman (2014) and implemented in the R-STAN statistical package. This is open-source software, with both R and STAN available in versions that run on Linux, Windows, and Mac platforms. The key acceptance stage of the sampler utilizes a Metropolis-Hastings algorithm, and the sampler incorporates an adaptive adjustment of Hamiltonian discrete time step size, and the interactive determination of number of steps to optimize acceptance rates. The sampler is flexible in specification of priors and likelihoods, and is not limited to the multivariate normal model described above. A drawback of this flexibility is that STAN may run more slowly for this model than procedures that use computationally efficient Gibbs samplers for part of the multivariate normal iteration. Suppose an intermediate parameter vector \( \nu \) that is multivariate normal with mean \( \mu \) and a covariance matrix \( \Omega = \mathbf{D} \mathbf{\Gamma} \mathbf{D}' \), where \( \mathbf{D} \) is diagonal and \( \mathbf{\Gamma} \) is the lower triangular Cholesky factor of the correlation matrix (so that \( \mathbf{\Gamma} \) is characteristically a lower triangular matrix with rows whose sums of squared elements are one). Recommended priors are a relatively diffuse multivariate normal prior for \( \mu \), relatively diffuse exponential densities for the elements of \( \mathbf{D} \), and a LKJ(1) density for the correlation matrix \( \mathbf{\Gamma}' \); see Lewandowski, Kurowicka, and Joe (2009).\(^{27}\)

---

\(^{26}\) If one wants a model with full intra-consumer as well as inter-consumer heterogeneity, the only change in the HB procedure described is to introduce an additional layer of multivariate normal intermediate parameters \( \nu_{mn} \) that vary by menu as well as subject and have a density centered at \( \nu_n \) with a covariance matrix adjusted in an additional Gibbs layer.

\(^{27}\) All correlation matrices can be written as \( \mathbf{\Gamma}' \), where \( \mathbf{\Gamma} \) is a lower triangular matrix with the property that the squares of the elements in each row sum to one. Then the elements \( (\Gamma_1, \ldots, \Gamma_i) \) are points in a unit sphere of dimension \( i \), and the method of Muller (1957) and Marsaglia (1972) can be used to draw points uniformly from this sphere; i.e., draw \( i \) iid standard normal deviates, and scale the vector of these deviates by dividing by the square root of their sum of squares.
**Terminal Steps:** After a few thousand “burning in” iterations $I_1$, consider the sampler data $\mu_i$, $\Omega_i$, and $v_{ni}$ over subjects $n = 1, \ldots, N$ and iterations $i = I_1 + 1, \ldots, I_1 + I$, where $I$ is typically taken to be at least 10,000, and perhaps as large as 1 million. The Bayes estimates of the deep parameters are the posterior means $\hat{\mu}$ and $\hat{\Omega}$ formed by averaging $\mu_i$ and $\Omega_i$ respectively over $i = 1, \ldots, I$. A first question in applications of sampling schemes like the one described above is whether one has achieved convergence to a stationary distribution that spans $\theta$-space, with sufficiently strong mixing so that serial correlations between distant iterations are small and sampler means are accurate estimates of true posterior means. The recommendations of the developers of these samplers are to start them from alternative parameter values, vary burn-in times and tuning parameters, and jitter tuning parameters to avoid accidental cyclicity. Econometric time series diagnostic tests for stationarity and for mixing conditions can be applied to the sampler output; see Azizzadeh and Rezakhah (2014), Delgado (2005), Domowitz and El-Gamal (1993), Corradi et al. (2000); Ljung and Box (1978). In the special case of test data with known deep parameters, it is straightforward to examine bias and mean square error as a function of sampler characteristics.28

From a classical perspective, the posterior means $\hat{\mu}$ and $\hat{\Omega}$ are point estimates of true parameter values $\mu_0$ and $\Omega_0$. Suppose the analyst wishes to treat these as classical estimators and carry out conventional steps such as associating standard errors with the estimates and providing confidence bounds or hypothesis tests. Due to the asymptotic equivalence of MLE and hierarchical Bayes estimators, this can be done using the formulas (13)-(17). To improve the asymptotic approximation, we recommend executing one or more BHHH steps starting from the HB posterior means, using the result from (15) to obtain MLE-equivalent estimates, and using (16) to estimate the covariance matrix of the resulting estimators. In most cases, it is practical to draw new intermediate parameter vectors $v_r$ for (13)-(17), but it is also possible to reuse the array of vectors $v_n$ from iterations in the HB process in the calculation of (15)-(17).

5.4. Policy Simulations. Once the Bayesian parameter estimates of $\mu$ and $\Omega$ are in hand, labeled $\hat{\mu}$ and $\hat{\Omega}$, consider hypothetical policy alternatives $m = 1, \ldots, M$, and let $(x_{mn}, q_{mn}, p_{mn})$ denote the economic environment of consumer decision-making under each policy. Note that this is a different use of the index $m$ than in the CBC where $m$ indexed menus presented to each subject. Take $R$ draws, labeled $r=1, \ldots, R$ of $v$ from $N(\hat{\mu}, \hat{\Omega})$ and calculate the corresponding taste parameters for each draw. Calculate

---

28 In test data where the vectors $v_n$ that determine individual taste parameters for each subject $n$ are drawn from a known normal distribution $N(\mu, \Omega)$, and choice data are then generated from a random utility model with these taste parameters the deep parameters that should be used for these comparisons are $\mu_0 = \frac{1}{N} \sum_{n=1}^{N} v_n$ and $\Omega_0 = \frac{1}{N} \sum_{n=1}^{N} (v_n - \mu_0)(v_n - \mu_0)'$ rather than $\mu$ and $\Omega$; as the target of the analysis is recovery of characteristics of realized tastes, and the analysis cannot identify the process that generated the realized tastes.
\[ P_{jmn} = \frac{1}{R} \sum_r \frac{\exp(V_{jmn(r)}^{\nu_r})}{1 + \Sigma_{i=1}^{r} \exp(V_{imn(r)}^{\nu_r})}. \]

From the \( P_{jmn} \) for each \( n \), one can estimate market shares (demands) for each product under each policy alternative, and if it is a target of the investigation, the policy impacts on selected sub-populations. The estimated probabilities (38) can also be used in (7) to estimate consumer welfare effects of policy changes. Confidence intervals can be obtained by implementing a Krinsky-Robb (1986, 1990) procedure with the estimated covariance of \( \hat{\mu} \) and \( \hat{\Omega} \), or by a block bootstrap in which repeated estimates of the deep parameters are obtained by resampling from the sample of survey respondents.29

5.5. A Monte Carlo Example. To demonstrate the procedures given in the previous two sections, consider a Monte Carlo example where a known decision-making process is used to generate an artificial CBC data set, and the methods above are then applied to estimate the parameters behind these “observed” choices. For this exercise, suppose we have CBC data on preferences for table grapes, the example discussed earlier in section 3.6. Suppose 1000 subjects are each presented with eight menus, and on each menu they have the option of choosing one of three bunches of grapes with various attributes and prices, or choosing to buy no grapes on this occasion. Assume that the subjects are prompted to think of this as the opportunities they will face on their next trip to the supermarket, so that existing stocks of grapes and purchasing opportunities outside the experiment are similar to those faced by these subjects in the real market. The product attributes and levels in the experiment are given in Table 4.

Table 4. Table Grape CBC Attributes and Levels

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Symbol</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>P</td>
<td>$1.00 to $4.00</td>
</tr>
<tr>
<td>Sweetness</td>
<td>S</td>
<td>Sweet (1) or Tart (0)</td>
</tr>
<tr>
<td>Crispness</td>
<td>C</td>
<td>Crisp (1) or Soft (0)</td>
</tr>
<tr>
<td>Size</td>
<td>L</td>
<td>Large (1) or Small (0)</td>
</tr>
<tr>
<td>Organic</td>
<td>O</td>
<td>Organic (1) or Non-organic (0)</td>
</tr>
</tbody>
</table>

The gender (G) of each subject is observed, coded \( G = 1 \) for female and \( G = 0 \) for male. Let \( I_n \) denote disposable income, which is not measured, but which drops out of the utility maximization. Define the interactions \( SC \equiv S*C \) and \( SG \equiv S*G \). Index a menu’s alternatives by \( j = 1, 2, 3 \) for the three offered bunches, and \( j = 4 \) for the no purchase alternative.

29 When the number of menus offered to each subject in the CBC experiment is finite, then inserting the mean of the posterior of the person-specific taste parameters into (32) provides a biased estimate of the subject’s probability.
option. Let $q_{jmn}$ be a dummy variable that is one for purchase alternatives ($j = 1, 2, 3$) and zero for $j = 4$. Assume the true utility has the WTP form

$$U_{jmn} = I_n - P_{jmn} + S_{jmn} \beta_s + C_{jmn} \beta_C + L_{jmn} \beta_L + O_{jmn} \beta_O + SC_{jmn} \beta_{SC} + SF_{jmn} \beta_{SF} + q_{jmn} \delta_n + \alpha_n \epsilon_{jmn},$$

where the $\epsilon_{jmn}$ are i.i.d. standard EV1 distributed, and $\nu_n = (\beta_s, \beta_C, \beta_L, \beta_O, \beta_{SC}, \beta_{SF}, \delta_n, \log(\alpha_n))$ are the preference parameters. Define $D_{jmn} = 1$ for the alternative within a menu that maximizes utility, $D_{jmn} = 0$ otherwise. Assume that the true parameters are distributed multivariate normal in the population with the means, standard deviations, and correlation/covariance matrix given in Table 5; Appendix C gives the R code that generates the data.

### Table 5. True Parameter Distribution $\beta_s, \beta_C, \beta_L, \beta_O, \beta_{SC}, \beta_{SF}, \delta_n, \log(\alpha_n)$

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Covariance (bottom) and Correlation (top) Matrices</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_s$</td>
<td>1.0</td>
<td>0.3</td>
<td>$\begin{bmatrix} 0.09 &amp; 0.6 &amp; 0 &amp; 0 &amp; 0.3 &amp; 0 &amp; 0 \end{bmatrix}$</td>
</tr>
<tr>
<td>$\beta_C$</td>
<td>0.3</td>
<td>0.1</td>
<td>$\begin{bmatrix} 0.018 &amp; 0.01 &amp; 0.48 &amp; 0 &amp; 0.42 &amp; 0 &amp; 0 \end{bmatrix}$</td>
</tr>
<tr>
<td>$\beta_L$</td>
<td>0.2</td>
<td>0.1</td>
<td>$\begin{bmatrix} 0 &amp; 0.0048 &amp; 0.01 &amp; 0 &amp; 0.42 &amp; 0 &amp; 0 \end{bmatrix}$</td>
</tr>
<tr>
<td>$\beta_O$</td>
<td>0.1</td>
<td>0.2</td>
<td>$\begin{bmatrix} 0 &amp; 0 &amp; 0 &amp; 0.04 &amp; 0.3 &amp; 0 &amp; 0 \end{bmatrix}$</td>
</tr>
<tr>
<td>$\beta_{SC}$</td>
<td>0.0</td>
<td>0.05</td>
<td>$\begin{bmatrix} 0.0045 &amp; 0.0021 &amp; 0.0021 &amp; 0.003 &amp; 0.0025 &amp; 0 &amp; 0 \end{bmatrix}$</td>
</tr>
<tr>
<td>$\beta_{SF}$</td>
<td>0.1</td>
<td>0.2</td>
<td>$\begin{bmatrix} 0 &amp; 0 &amp; 0 &amp; 0 &amp; 0.04 &amp; 0 &amp; 0 \end{bmatrix}$</td>
</tr>
<tr>
<td>$\delta_n$</td>
<td>2.0</td>
<td>1.0</td>
<td>$\begin{bmatrix} 0 &amp; 0 &amp; 0 &amp; 0 &amp; 0 &amp; 0 &amp; 1 \end{bmatrix}$</td>
</tr>
<tr>
<td>$\log(\alpha_n)$</td>
<td>-0.5</td>
<td>0.3</td>
<td>$\begin{bmatrix} 0 &amp; 0 &amp; 0 &amp; 0 &amp; 0 &amp; 0 &amp; 0.09 \end{bmatrix}$</td>
</tr>
</tbody>
</table>

The model (39) is estimated with independence of $\log(\alpha_n)$ and the remaining parameters imposed, but all other covariances allowed. Table 6ABC gives estimates by hierarchical Bayes using the Allenby-Train procedure (HB-AT), hierarchical Bayes using the NUTS procedure (HB-NUTS), and maximum simulated likelihood (MSL). Note that the true population parameter values given in Table 4 are reproduced closely but not exactly by their sample counterparts for the 1000 subjects in the CBC study; these are given in the third column of the table, and are the relevant values for comparison with estimates. The maximum simulated likelihood estimates use 250 Halton draws to approximate the terms in (15) and (16), and uses Newton-Raphson iteration to the estimator. Starting values were obtained by estimating a model with fixed coefficients, using the estimated WTPs as starting values in a model with random but uncorrelated WTP’s, and then using these estimates as starting values in the model with correlated WTP’s. The three steps combined took about three hours of run time. The HB-AT estimates are obtained with 100,000 iterations used to tune and burn-in the sampler, and every tenth draw taken from 100,000 subsequent draws.
to form the posterior means given in the table. It has been observed that this procedure tends to inflate variances that are numerically very small (see, e.g., Balcombe et al, 2009); to counteract this tendency, we divided the nonprice variables by ten and then rescaled the estimates accordingly. Estimation took 12 minutes. The HB-NUTS estimates are obtained from 11,000 iterations, with the first 5500 discarded for “burn-in”, and every 10th draw from the remainder used to obtain the estimates in Table 6.

Table 6A. Estimates from Table Grapes CBC Data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Values</th>
<th>HB-AT</th>
<th>HB-NUTS</th>
<th>MSL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Population</td>
<td>Sample Mean</td>
<td>Std Dev</td>
<td>Mean</td>
</tr>
<tr>
<td>Means</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_S$</td>
<td>1.0</td>
<td>1.005</td>
<td>0.9682</td>
<td>(0.0445)</td>
</tr>
<tr>
<td>$\beta_C$</td>
<td>0.3</td>
<td>0.299</td>
<td>0.3495</td>
<td>(0.0358)</td>
</tr>
<tr>
<td>$\beta_L$</td>
<td>0.2</td>
<td>0.197</td>
<td>0.1781</td>
<td>(0.0251)</td>
</tr>
<tr>
<td>$\beta_O$</td>
<td>0.1</td>
<td>0.109</td>
<td>0.0910</td>
<td>(0.0238)</td>
</tr>
<tr>
<td>$\beta_{SC}$</td>
<td>0.0</td>
<td>0.001</td>
<td>-0.0450</td>
<td>(0.0485)</td>
</tr>
<tr>
<td>$\beta_{SG}$</td>
<td>0.1</td>
<td>0.093</td>
<td>0.1369</td>
<td>(0.0540)</td>
</tr>
<tr>
<td>$\delta_n$</td>
<td>2.0</td>
<td>2.012</td>
<td>2.0049</td>
<td>(0.0498)</td>
</tr>
<tr>
<td>$\log(\alpha_n)$</td>
<td>-0.5</td>
<td>-0.481</td>
<td>-0.5277</td>
<td>(0.0253)</td>
</tr>
</tbody>
</table>

Table 6B. Estimates from Table Grapes CBC Data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Values</th>
<th>HB-AT</th>
<th>HB-NUTS</th>
<th>MSL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Population</td>
<td>Sample Mean</td>
<td>Std Dev</td>
<td>Mean</td>
</tr>
<tr>
<td>Std Devs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_S$</td>
<td>0.3</td>
<td>0.303</td>
<td>0.3086</td>
<td>(0.0633)</td>
</tr>
<tr>
<td>$\beta_C$</td>
<td>0.1</td>
<td>0.099</td>
<td>0.2149</td>
<td>(0.0354)</td>
</tr>
<tr>
<td>$\beta_L$</td>
<td>0.1</td>
<td>0.101</td>
<td>0.1621</td>
<td>(0.0370)</td>
</tr>
<tr>
<td>$\beta_O$</td>
<td>0.2</td>
<td>0.210</td>
<td>0.2157</td>
<td>(0.0378)</td>
</tr>
<tr>
<td>$\beta_{SC}$</td>
<td>0.05</td>
<td>0.051</td>
<td>0.2227</td>
<td>(0.0516)</td>
</tr>
<tr>
<td>$\beta_{SG}$</td>
<td>0.2</td>
<td>0.202</td>
<td>0.2518</td>
<td>(0.0758)</td>
</tr>
<tr>
<td>$\delta_n$</td>
<td>1.0</td>
<td>0.988</td>
<td>0.7924</td>
<td>(0.0764)</td>
</tr>
<tr>
<td>$\log(\alpha_n)$</td>
<td>0.3</td>
<td>0.296</td>
<td>0.4220</td>
<td>(0.0257)</td>
</tr>
</tbody>
</table>

30 We set T, the degrees of freedom for the inverted Wishart prior, equal to the number of random coefficients in the model. We have found little numerical effect of using a larger degrees of freedom.
### Table 6C. Estimates from Table Grapes CBC Data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Values</th>
<th>HB-AT</th>
<th>HB-NUTS</th>
<th>MSL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Population</td>
<td>Sample</td>
<td>Mean</td>
<td>Std Dev</td>
</tr>
<tr>
<td>Covariances</td>
<td></td>
<td></td>
<td>Mean</td>
<td>Std Dev</td>
</tr>
<tr>
<td>$\beta_s, \beta_c$</td>
<td>0.018</td>
<td>0.0185</td>
<td>0.0348</td>
<td>0.0183</td>
</tr>
<tr>
<td>$\beta_s, \beta_L$</td>
<td>0</td>
<td>0.0007</td>
<td>-0.0054</td>
<td>0.0166</td>
</tr>
<tr>
<td>$\beta_s, \beta_O$</td>
<td>0</td>
<td>-0.0001</td>
<td>0.0077</td>
<td>0.0169</td>
</tr>
<tr>
<td>$\beta_s, \beta_{SC}$</td>
<td>0.0045</td>
<td>0.0046</td>
<td>-0.0244</td>
<td>0.0237</td>
</tr>
<tr>
<td>$\beta_s, \beta_{SG}$</td>
<td>0</td>
<td>-0.0004</td>
<td>0.0002</td>
<td>0.0230</td>
</tr>
<tr>
<td>$\beta_s, \delta_n$</td>
<td>0</td>
<td>-0.0032</td>
<td>0.1430</td>
<td>0.0566</td>
</tr>
<tr>
<td>$\beta_c, \beta_L$</td>
<td>0.0048</td>
<td>0.0050</td>
<td>0.0067</td>
<td>0.0095</td>
</tr>
<tr>
<td>$\beta_c, \beta_O$</td>
<td>0</td>
<td>0.0006</td>
<td>0.0161</td>
<td>0.0119</td>
</tr>
<tr>
<td>$\beta_c, \beta_{SC}$</td>
<td>0.0021</td>
<td>0.0022</td>
<td>-0.0291</td>
<td>0.0154</td>
</tr>
<tr>
<td>$\beta_c, \beta_{SG}$</td>
<td>0</td>
<td>0.0000</td>
<td>-0.0129</td>
<td>0.0177</td>
</tr>
<tr>
<td>$\beta_c, \delta_n$</td>
<td>0</td>
<td>-0.0027</td>
<td>0.1073</td>
<td>0.0402</td>
</tr>
<tr>
<td>$\beta_O, \beta_c$</td>
<td>0</td>
<td>0.0009</td>
<td>0.0145</td>
<td>0.0114</td>
</tr>
<tr>
<td>$\beta_O, \beta_{SC}$</td>
<td>0.0021</td>
<td>0.0023</td>
<td>-0.0052</td>
<td>0.0116</td>
</tr>
<tr>
<td>$\beta_O, \beta_{SG}$</td>
<td>0</td>
<td>0.0001</td>
<td>-0.0105</td>
<td>0.0148</td>
</tr>
<tr>
<td>$\beta_O, \delta_n$</td>
<td>0</td>
<td>-0.0040</td>
<td>-0.0013</td>
<td>0.0332</td>
</tr>
<tr>
<td>$\beta_{SC}, \beta_{SG}$</td>
<td>0.0030</td>
<td>0.0032</td>
<td>-0.0142</td>
<td>0.0131</td>
</tr>
<tr>
<td>$\beta_{SC}, \delta_n$</td>
<td>0</td>
<td>-0.0016</td>
<td>-0.0153</td>
<td>0.0202</td>
</tr>
<tr>
<td>$\beta_{SG}, \delta_n$</td>
<td>0</td>
<td>-0.0002</td>
<td>0.0492</td>
<td>0.0331</td>
</tr>
<tr>
<td>$\beta_{SC}, \beta_{SG}$</td>
<td>0</td>
<td>-0.0002</td>
<td>0.0212</td>
<td>0.0194</td>
</tr>
<tr>
<td>$\beta_{SC}, \delta_n$</td>
<td>0</td>
<td>-0.0020</td>
<td>-0.1083</td>
<td>0.0513</td>
</tr>
<tr>
<td>$\beta_{SG}, \delta_n$</td>
<td>0</td>
<td>0.0008</td>
<td>-0.0417</td>
<td>0.0615</td>
</tr>
</tbody>
</table>

For the NUTS sampler, we also calculated posterior means (not reported) using all iterations including burn-in iterations, using every iteration rather than every 10th iteration, and using every 100th iteration including burn-in iterations. The results were very close, indicating that in this application NUTS seems to show rapid convergence with relatively little drift. However, time-series graphs do show some long cycles in the parameter estimates. The NUTS procedure required computation overnight; this may be due to storage of posterior intermediate parameters for each subject in each iteration, output that may be useful for some marketing applications such as segmentation, but is not needed for the primary task of estimating the deep parameters. Comparing the HB-AT, HB-NUTS, and
MSL estimates with the true parameter values and with each other, we find that all the methods do a good job of recovering the true utility model.

6. AN EMPIRICAL CBC STUDY USING MSL AND HB METHODS

In this section, we provide an example of a money-metric random utility model estimated using both Maximum Simulated Likelihood and Hierarchical Bayes methods. We utilize the data from a conjoint experiment designed and described by Butler and Glasgow (2015) for consumers’ choice among video streaming services. Their experiments included the monthly price of the service and the following non-price attributes:

Table 7. Non-Price Attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commercials shown between content</td>
<td>Yes (‘‘commercials’’)</td>
</tr>
<tr>
<td></td>
<td>No (baseline category)</td>
</tr>
<tr>
<td>Speed of content availability</td>
<td>TV episodes next day, movies in 3 months (‘‘fast content’’)</td>
</tr>
<tr>
<td></td>
<td>TV episodes in 3 months, movies in 6 months (baseline category)</td>
</tr>
<tr>
<td>Catalogue</td>
<td>10,000 movies and 5,000 TV episodes (‘‘mostly movies’’)</td>
</tr>
<tr>
<td></td>
<td>2,000 movies and 13,000 TV episodes (‘‘mostly TV shows’’)</td>
</tr>
<tr>
<td></td>
<td>5,000 movies and 2,500 TV episodes (the baseline category)</td>
</tr>
<tr>
<td>Data-sharing policies</td>
<td>Information is collected but not shared (baseline category)</td>
</tr>
<tr>
<td></td>
<td>Usage information is share with third parties (‘‘share usage’’)</td>
</tr>
<tr>
<td></td>
<td>Usage and personal information are shared with third parties (‘‘share usage and personal’’)</td>
</tr>
</tbody>
</table>

Each choice experiment included four alternative video streaming services with specified price and attributes plus a fifth alternative of not subscribing to any video streaming service. Each respondent was presented with 11 choice experiments. Butler and Glasgow obtained choice from 300 respondents and implemented an estimation procedure that accounts for “protestors” (mainly consumers who never chose a service that shared data.) In our use of their data, we do not include the 40 respondents that they identified as protestors, so that our sample consists of 260 respondents.

Table 8 gives a model estimated in WTP-space by hierarchical Bayes, using the Allenby-Train procedure described above with 10,000 burn-in iterations, 10,000 iterations after burn-in from which, to reduce serial correlation, every 10th draw from the sampler was retained to calculate the estimates. WTP for each non-price attribute is assumed to be normally distributed over consumers; the price/scaling coefficient (1/α) is assumed to be log-normally distributed; and correlation is allowed among the WTP’s as well as the price/scaling coefficient. The point estimate of the population mean in the mean of the draws from the posterior distribution, and the standard error of this estimate is the standard deviation of the draws. Similarly for the standard deviation in the population:
the estimate is the mean of the draws from its posterior and the standard error is the standard deviation of these draws. The point estimates of the correlations are shown in the second part of the table.

**Table 8A. Hierarchical Bayes Model of WTPs for Video Streaming Service**

<table>
<thead>
<tr>
<th></th>
<th>Population Mean</th>
<th>Std Dev in Population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Std Error</td>
</tr>
<tr>
<td>Log(1/α)</td>
<td>-1.890</td>
<td>0.1188</td>
</tr>
<tr>
<td>WTP for:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commercials</td>
<td>-1.571</td>
<td>0.5547</td>
</tr>
<tr>
<td>Fast Availability</td>
<td>3.496</td>
<td>0.7182</td>
</tr>
<tr>
<td>Mostly TV shows</td>
<td>0.227</td>
<td>0.5237</td>
</tr>
<tr>
<td>Mostly movies</td>
<td>2.569</td>
<td>0.5170</td>
</tr>
<tr>
<td>Share usage only</td>
<td>-0.432</td>
<td>0.3015</td>
</tr>
<tr>
<td>Share personal and usage</td>
<td>-1.831</td>
<td>0.6325</td>
</tr>
<tr>
<td>No service</td>
<td>-28.591</td>
<td>3.2753</td>
</tr>
</tbody>
</table>

**Table 8B. Correlations: point estimates**

<table>
<thead>
<tr>
<th></th>
<th>Log(1/α)</th>
<th>Commercials</th>
<th>Fast Availability</th>
<th>Mostly TV</th>
<th>Mostly movies</th>
<th>Share usage</th>
<th>Share personal and usage</th>
<th>No service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log(1/α)</td>
<td>1.0000</td>
<td>-0.3010</td>
<td>-0.1484</td>
<td>-0.1414</td>
<td>0.0072</td>
<td>-0.0050</td>
<td>-0.1431</td>
<td>0.3865</td>
</tr>
<tr>
<td>Commercials</td>
<td>1.0000</td>
<td>0.2238</td>
<td>0.1493</td>
<td>-0.1912</td>
<td>-0.0799</td>
<td>-0.0204</td>
<td>-0.5369</td>
<td></td>
</tr>
<tr>
<td>Fast Availability</td>
<td>1.0000</td>
<td>0.5094</td>
<td>0.0418</td>
<td>-0.5126</td>
<td>-0.5135</td>
<td>0.2197</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mostly TV</td>
<td>1.0000</td>
<td>-0.4523</td>
<td>-0.1151</td>
<td>-0.2787</td>
<td>0.3266</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mostly movies</td>
<td>1.0000</td>
<td>-0.0057</td>
<td>0.1650</td>
<td>0.1884</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Share usage</td>
<td>1.0000</td>
<td>0.8155</td>
<td>-0.2866</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Share personal and usage</td>
<td>1.0000</td>
<td>-0.4535</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No service</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results in Table 8 indicate that people are willing to pay $1.57 per month, on average, to avoid having commercials before their content. Fast availability is valued highly, with an average WTP of $3.50 per month in order to see TV shows and movies soon after their original showing. Some people prefer having mostly TV shows rather than a more even mix of TV and movies; but more people prefer having mostly movies over the mix. The
correlation between these two coefficients is -0.45, showing, as expected, that people who prefer mostly TV over a mix tend to prefer a mix over having mostly movies. Interestingly, people who want fast availability tend to be those who prefer mostly TV shows: the correlation between these two WTP’s is 0.51, while the correlation between WTP for fast availability and mostly movies is only 0.04. Apparently, the desire for fast availability mainly applies to TV shows.

Consider now the data-sharing policies of services. The point estimate implies that consumers are a WTP of 43 cents per month to avoid having their usage data shared in aggregate form; however, the hypothesis of zero average WTP cannot be rejected. Consumers are much more concerned about their personal information being shared along with their usage information: the average WTP to avoid such sharing is $1.83 per month. The correlation between WTP to avoid the two forms of sharing is a substantial 0.82.

Table 9 gives the same model estimated by the method of maximum simulated likelihood (MSL). We used STATA’s module for mixed logit in wtp-space, “mixlogitwtp.” We specified 100 Halton draws per person, which has been shown to be more accurate for mixed logit estimation than 1000 pseudo-random draws per person (Bhat, 2001; Train, 2000, 2009). We used the HB estimates (Table 7) as starting values. At the HB estimates, the log-likelihood was -3994.64 and rose to -3903.47 at convergence.32

The MSL estimates for the mean WTP’s are fairly similar to the HB estimates. In particular, the MSL estimates of mean WTP to: avoid commercials is $1.56 compared to $1.57 by HB; obtain fast availability is $3.95 compared to $3.50 by HB; obtain the mix with more movies is $2.96 compared to $2.57 by HB; to avoid having no service is $27.26 compared to $28.59 by HB. The mean WTP’s for the mix with more TV shows and to avoid sharing of usage data are insignificant under both MSL and HB. The only substantial difference is the mean WTP to avoid sharing of personal and usage information, which is estimated to be $2.71 by MSL and $1.83 by HB.

The estimated standard deviations of WTP are practically the same by MSL and HB for the following attributes: to avoid commercials, obtain fast service, obtain a mix with mostly movies, and avoid not having service. Standard deviations of WTP for the mix with mostly TV shows and to avoid either form of data sharing are estimated to be higher by MSL than HB. The correlations that were discussed above in relation to the HB estimates are also evidenced with MSL. However, other correlations differ in magnitude and even sign between MSL and HB.

31 The module, created by Prof. Arne Rise Hole of the University of Sheffield, is available at https://www.sheffield.ac.uk/economics/people/hole/stata or by typing “ssc install mixlogitwtp” at the STATA command prompt. For mixed logit models in preference space, his module “mixlogit” is available. The two modules share syntax and options.

32 The mixlogitwtp module allows estimation by Newton-Raphson (NR), BHHH, DFP, and BFGS. Using NR, one iteration raised the log-likelihood from -3994.65 to -3969.31. The remaining rise to -3903.47 at convergence took 29 iterations. Run time was about 90 minutes. The NR option utilizes numerical gradients and Hessian. The other procedures utilize numerical gradients but alternative forms of the Hessian and can be expected to run faster.
Table 9A. MSL Estimates of WTPs for Video Streaming Services

<table>
<thead>
<tr>
<th></th>
<th>Population Mean</th>
<th>Std Dev in Population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Std Error</td>
</tr>
<tr>
<td>Log(1/α)</td>
<td>-2.002</td>
<td>0.945</td>
</tr>
<tr>
<td>WTP for:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commercials</td>
<td>-1.562</td>
<td>0.4214</td>
</tr>
<tr>
<td>Fast Availability</td>
<td>3.945</td>
<td>0.4767</td>
</tr>
<tr>
<td>Mostly TV shows</td>
<td>-0.6988</td>
<td>0.4783</td>
</tr>
<tr>
<td>Mostly movies</td>
<td>2.963</td>
<td>0.4708</td>
</tr>
<tr>
<td>Share usage only</td>
<td>-0.6224</td>
<td>0.4040</td>
</tr>
<tr>
<td>Share personal and usage</td>
<td>-2.705</td>
<td>0.5844</td>
</tr>
<tr>
<td>No service</td>
<td>-27.26</td>
<td>2.662</td>
</tr>
</tbody>
</table>

Table 9B. Correlations: point estimates

(* denotes significance at 5% level)

<table>
<thead>
<tr>
<th></th>
<th>Log(1/α)</th>
<th>Commercials</th>
<th>Fast Availability</th>
<th>Mostly TV</th>
<th>Mostly movies</th>
<th>Share usage</th>
<th>Share personal and usage</th>
<th>No service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log(1/α)</td>
<td>1.0000</td>
<td>-0.5813*</td>
<td>-0.1371</td>
<td>0.0358</td>
<td>0.0256</td>
<td>0.0022</td>
<td>-0.1287</td>
<td>0.2801*</td>
</tr>
<tr>
<td>Commercials</td>
<td>1.0000</td>
<td>-0.3473*</td>
<td>-0.4019*</td>
<td>0.0109</td>
<td>-0.2562</td>
<td>-0.0079</td>
<td>-0.4206*</td>
<td>-0.4108*</td>
</tr>
<tr>
<td>Fast Availability</td>
<td>1.0000</td>
<td>0.8042*</td>
<td>-0.3542*</td>
<td>-0.3542*</td>
<td>-0.4206*</td>
<td>-0.4206*</td>
<td>0.2391*</td>
<td></td>
</tr>
<tr>
<td>Mostly TV</td>
<td>1.0000</td>
<td>-0.5890*</td>
<td>-0.1695</td>
<td>-0.3328*</td>
<td>0.4616*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mostly movies</td>
<td>1.0000</td>
<td>0.5141*</td>
<td>0.5181*</td>
<td>-0.0147</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Share usage</td>
<td>1.0000</td>
<td>0.9370*</td>
<td>0.0975</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Share personal and usage</td>
<td>1.0000</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No service</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.0000</td>
</tr>
</tbody>
</table>

A model without the money-metric scaling, called a preference space model, was also estimated (not shown.) In general, the estimated means and standard deviations of WTP were considerably higher and less plausible. For example, for sharing of usage and personal data, the model in WTP-space above gives a mean WTP to avoid sharing of $2.71 with a standard deviation of $6.75. For the model in preference space, the derived distribution of WTP has
a mean $6.99 and standard deviation of $98.95. The high values, especially for the standard deviation, are due to the fact that WTP in the preference space model is calculated as the ratio of the attribute’s coefficient divided by the price coefficient, such that values of the price coefficient that are arbitrarily close to zero (which are allowed by the distribution of the price coefficient) imply arbitrarily large WTP. Models in WTP-space avoid this issue by specifying and estimating the distribution of WTP directly. However, the log-likelihood of the model in preference space is higher than that in wtp-space: -3863.9 compared to -3903.5. These results, namely, more plausible estimates of the distribution of WTP accompanied by a lower log-likelihood, mirror those of Weeks and Train (2005). Other studies making this comparison (e.g., Scarpa et al. 2008) have found no such tradeoff on their datasets, with their model in WTP-space obtaining a better fit as well as more plausible distributions of WTP.

As discussed above, the means and standard deviations of the WTP estimates obtained by MSL and HB-AT methods generally agree on sign and magnitude, but are not close enough to conclude that these methods have converged to asymptotically equivalent estimates. This may reflect finite sampling deviations from the asymptotic limit; there is no guarantee that the hierarchical Bayes and MSL estimates will be close in finite samples. However, both methods use simulations, and the hierarchical Bayes methods in particular are known to converge slowly to their posterior limiting distribution when their tuning is imperfect. Then, some of the differences in Tables 7 and 8 might disappear with more complete tuning and more simulation draws.

7. CONCLUSIONS

This paper has reviewed methods for stated preference elicitation, concentrating on the most successful and widely applied approach, choice-based conjoint analysis (CBC). It has reviewed the choice-theory underpinnings of CBC, and recommends modeling utility in money-metric or WTP-space form, introducing flexible generic attributes, controlling carefully how components of income enter utility, discrete choice, and calculations of consumer welfare. It has reviewed the design of CBC experiments, and outlined the conditions under which such experiments seem to be able to reliably mimic consumers’ market experience and predict their market behavior. It has discussed extensively the use of simulated maximum likelihood and hierarchical Bayes methods to estimate the deep parameters that characterize distributions of tastes and choices in the population.

Our overall conclusion is that collection and analysis of stated preference data is a powerful tool for exploring and predicting consumer behavior, but both the data collection and analysis require considerable care and caution in inferring that the results are predictive and reliable for real market behavior. In applications such as demand for complex, unfamiliar products, or products that have substantial public goods aspects such as remedies for non-use environmental damages, where there are no good market benchmarks, we conclude that the reliability of stated preference methods has not been established.
REFERENCES


Koop, G. (2003), *Bayesian Econometrics*, John Wiley & Sons Ltd.


Rossi, R.; G. Allenby; R. McCulloch (2005), Bayesian Statistics and Marketing, John Wiley & Sons Ltd.


Web Sites:
https://github.com/stan-dev/rstan/wiki/RStan-Getting-Started
http://mc-stan.org/
http://www.slideshare.net/surveyanalytics/how-to-run-a-conjoint-analysis-project-in-1-hour
http://cran.r-project.org/web/packages/conjoint/index.html
APPENDIX

A. Stan code for the hierarchical models in the paper (hb.stan)

// Note: Number of alternatives (4) hard coded

data {
  int<lower=0> K_u;            // Number of variables with uncorrelated coefs.
  int<lower=0> K_c;            // Number of variables with correlated coefs.
  int<lower=1> R;              // Number of respondents
  int<lower=1> S;              // Number of choice sets per respondent
  int<lower=1> N;              // Total number of choice sets observed
  int<lower=1,upper=R> r[N];   // Respondent ID
  int<lower=1,upper=S> y[N];   // Alternative selected
  real x1_u[N];                // Variables with uncorrelated coefs for choice 1
  real x2_u[N];                // Variables with uncorrelated coefs for choice 2
  real x3_u[N];                // Variables with uncorrelated coefs for choice 3
  real x4_u[N];                // Variables with uncorrelated coefs for choice 4
  row_vector[K_c] x1_c[N];     // Variables with correlated coefs for choice 1
  row_vector[K_c] x2_c[N];     // Variables with correlated coefs for choice 2
  row_vector[K_c] x3_c[N];     // Variables with correlated coefs for choice 3
  row_vector[K_c] x4_c[N];     // Variables with correlated coefs for choice 4
}

parameters {
  vector[K_c] beta[R];                 // Correlated coefficients for each respondent
  vector[K_c] mu_c;                    // Means of correlated coefficients
  corr_matrix[K_c] Omega;              // Correlation matrix
  real<lower=0,upper=100> tau;         // Variances of correlated coefficients
  vector<lower=0,upper=100>[K_c] mu_p; // Mean of alpha
  vector<lower=0,upper=100> sigma_p;  // Variance of alpha
}

model {
  matrix[K_c,K_c] Sigma;  // Covariance matrix
  Sigma <- quad_form_diag(Omega,tau);  // Prior for tau flat on [0,100]
  // tau ~ cauchy(0,2.5); // Alternative prior for tau
  Omega ~ lkj_corr(1); // Prior for Omega (correlation matrix)
  // mu_c ~ normal(0, 100); // Prior for mean non-price coefs.
  mu_p ~ normal(0, 100); // Prior for mean alpha
  for (i in 1:R){ // Respondent-specific draws
    beta[i] ~ multi_normal(mu_c, Sigma); // Model for individual beta coefs.
    alpha[i] ~ lognormal(mu_p, sigma_p); // Model for individual alpha coefs.
  }

  for (n in 1:N){ // Choice probabilities
    vector[4] p;  // Choice probabilities
    p[1] <- exp(-x1_u[n] / alpha[r[n]] + x1_c[n] * beta[r[n]] / alpha[r[n]]);
    p[2] <- exp(-x2_u[n] / alpha[r[n]] + x2_c[n] * beta[r[n]] / alpha[r[n]]);
    p[3] <- exp(-x3_u[n] / alpha[r[n]] + x3_c[n] * beta[r[n]] / alpha[r[n]]);
    p[4] <- exp(-x4_u[n] / alpha[r[n]] + x4_c[n] * beta[r[n]] / alpha[r[n]]);
    p <- p / sum(p);
    y[n] ~ categorical(p); // Model for choice
  }
}

B. R code to implement Stan estimation
library(mlogit)
library(rstan)

### Load and organized data

data <- read.csv("data/grapes_data.csv")
data[, 1] <- NULL
names(data) <- c("participant", "set", "alternative", "price", "is_sweet", "is_crisp", "is_big", "is_organic", "is_sweet_crisp", "is_sweet_female", "is_grapes", "is_female", "is_chosen", "utility_actual", "utility_expected", "choice_prob")

### Ensure order for later calculations
data <- data[with(data, order(participant, set, alternative)), ]

### Hierarchical Bayesian Estimation

### Set up data and models

### Need unique identifiers for each choice situation for 'mlogit'
data$situation <- with(data, interaction(participant, set))
data_ml <- mlogit.data(data, shape = "long", choice = "is_chosen", alt.var = "alternative", chid.var = "situation", id.var = "participant")

### Select variables for model; 'is_female' not identified
variables <- c("price", grep("is_", names(data), value = TRUE))
variables <- variables[!variables %in% c("is_chosen", "is_female")]

### Create model formula
model <- paste("is_chosen", "~", paste(variables, collapse = " + "), "- 1")
model <- formula(model)

### HB tuning parameters
iter   <- 11000
chains <- 1

### Organize data for Stan

### Price coefficient not correlated with others
data_u <- model.matrix(~ price - 1, data = data)

### Remaining coefficients
data_c <- model.matrix(update(model, ~ . - price), data = data)

### Separate the covariates by alternative
x1_u <- data_u[data$alternative == 1L, ]
x2_u <- data_u[data$alternative == 2L, ]
```r
x3_u <- data_u[data$alternative == 3L, ]
x4_u <- data_u[data$alternative == 4L, ]

x1_c <- data_c[data$alternative == 1L, , drop = FALSE]
x2_c <- data_c[data$alternative == 2L, , drop = FALSE]
x3_c <- data_c[data$alternative == 3L, , drop = FALSE]
x4_c <- data_c[data$alternative == 4L, , drop = FALSE]

### Constants
### Assumes that all participants face same number of choice sets
K_u <- ncol(data_u)    # Number of uncorrelated variables
K_c <- ncol(data_c)    # Number of correlated variables
R   <- length(unique(data$participant)) # Number of participants
N   <- nrow(x1_c)      # Number of choice sets * participants
S   <- N/R             # Number of choice sets per person

y   <- data[data$is_chosen == 1L, "alternative"]   # Alternative selected for each set
r   <- data[data$alternative == 1L, "participant"] # Participant ID for each set

### Variables passed to Stan
vars_hb <- c(ls(pattern = "x\[[[:digit:]]\]_\{uc\}"), "K_u", "K_c", "R", "N", "S", "y", "r")

##############################
### Estimate model
##############################
model_hb <- stan(file = "hb.stan", data = vars_hb,
iter = iter, chains = chains)
save(model_hb, file = "results/hb-11000.RData")
```

C. R code for the table grape simulated CBC study

```r
N <- 1000 # Number of subjects
M <- 8    # Number of choice scenarios
# J <- 4  # Number of alternatives; Hard coded

# R session to create Monte Carlo example of table grape CBC
# indices
# subject (n)
# menu (m)
# alternative (j)
# factors
# price (P), in dollars/bunch from $1.00 to $4.00
# Sweetness (S), levels S = 0 for tart, S = 1 for sweet
# Chrispness (C), levels C = 0 for solt, C = 1 for crisp
# Size (L), levels L = 0 for small, L = 1 for large
# Organic (O), levels O = 1 if yes, O = 0 if no
# Sweetness*Crispness interaction (SC), levels S*C
# Sweetness*Gender interaction (SG), levels S*G, where G = 1 if female, G = 0 if male
# bunch-specific dummy (q)
# other variables
# Gender (G): G = 1 if female, G = 0 if male
# Choice (D): D = 1 if stated choice, D = 0 otherwise
# Realized utility (U)
# Expected utility (V)
# Choice probabilities (PR)
```

# the coefficient on price is normalized to -1
# taste parameters are for the remaining 7 factors, in the order above, plus the disturbance
# scale factor
parname <- c("bS", "bC", "bL", "bO", "bSC", "bSG", "bq", "lnalpha")  # parameter mnemonics
mu <- matrix(c(1.0, 0.3, 0.2, 0.1, 0.00, 0.1, 2.0, -0.5), 8, 1)  # true taste means
std <- matrix(c(0.3, 0.1, 0.1, 0.2, 0.05, 0.2, 1.0, 0.3), 8, 1)  # true taste standard deviations.
cho <- matrix(c(1.0, 0.6, 0.0, 0.0, 0.30, 0.0, 0.0, 0.0), 8, 1)  # true taste Cholesky factor correlation matrix
cho <- cbind(cho, matrix(c(0.0, 0.8, 0.6, 0.0, 0.30, 0.0, 0.0, 0.0), 8, 1))
cho <- cbind(cho, matrix(c(0.0, 0.0, 0.8, 0.0, 0.30, 0.0, 0.0, 0.0), 8, 1))
cho <- cbind(cho, matrix(c(0.0, 0.0, 0.0, 1.0, 0.30, 0.0, 0.0, 0.0), 8, 1))
cho <- cbind(cho, matrix(c(0.0, 0.0, 0.0, 0.0, 0.80, 0.0, 0.0, 0.0), 8, 1))
cho <- cbind(cho, matrix(c(0.0, 0.0, 0.0, 0.0, 0.00, 1.0, 0.0, 0.0), 8, 1))
cho <- cbind(cho, matrix(c(0.0, 0.0, 0.0, 0.0, 0.00, 0.0, 1.0, 0.0), 8, 1))
cho <- cbind(cho, matrix(c(0.0, 0.0, 0.0, 0.0, 0.00, 0.0, 0.0, 1.0), 8, 1))
theta0 <- cbind(mu, std, cho)  # true deep parameters

# initialize
coeff <- matrix(0, 8, N)  # array of subject taste coefficients
set.seed(1937)

# for (n in 1:N) {
#   coeff[n,] <- mu + std*cho * matrix(rnorm(8), 8, 1)  # individual true parameters
# }

dta <- function(i, k) {
  x <- matrix(0, 4, 16)
  x[,1] <- i
  x[,2] <- k
  x[1:3, 3] <- c(1, 2, 3, 4)
  x[1:3, 4] <- 1 + 3*round(runif(3), digits=2)  # P
  x[1:3, 5:8] <- ifelse(runif(12) > 0.5, 1, 0)  # S, C, L, O
  x[1:3, 9] <- x[1:3, 5]*x[1:3, 6]  # SC
  x[1:3, 10] <- g*x[1:3, 5]  # SG
  x[1:3, 11] <- 1  # q
  x[,12] <- g  # g
  x[1,15] <- -(x[,4] + x[,5:11])  # coeff[1:7, i])  # V
  x[1,14] <- x[1,15] + exp(coeff[8,i])*matrix(-log(-log(runif(4))), 4, 1)  # U
  u <- max(x[,14])
  x[,13] <- ifelse(x[,14] <= u, 1, 0)  # D
  s <- exp(x[,15]/exp(coeff[8,i]))
  x[,16] <- s/sum(s)  # PR
  return(x)
}

X <- NULL

# for (n in 1:N) {
#   g <- ifelse(runif(1) > 0.5, 1, 0)
#   for (m in 1:M) {
#     X <- rbind(X, dta(n, m))
#   }
# }

coeff <- t(coeff)
population <- data.frame(mu = colMeans(coeff), sd = apply(coeff, 2, sd))
population <- cbind(population, cor(coeff))
write.csv(X, file = "data/grapes_data.csv")
write.csv(varname, file = "data/grapes_data_names.csv")
write.csv(theta0, file = "results/theta0.csv")
write.csv(population, file = "results/population.csv")

The correlation matrix in Table 4 comes from the Cholesky factor matrix:

\[
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.6 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.6 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0.3 & 0.3 & 0.3 & 0.3 & 0.8 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
\end{pmatrix}
\]