

5-2010

# Reality Check: Combining Survey and Market Data to Estimate the Importance of Product Attributes

Elea M. Feit
*University of Pennsylvania*

Mark A. Beltramo

Fred M. Feinberg



Recommended Citation

Feit, E. M., Beltramo, M. A., & Feinberg, F. M. (2010). Reality Check: Combining Survey and Market Data to Estimate the Importance of Product Attributes. *Management Science, 56* (5), 785-800. http://dx.doi.org/10.1287/mnsc.1090.1136

# Reality Check: Combining Survey and Market Data to Estimate the Importance of Product Attributes

**Abstract**
Discrete choice models estimated using hypothetical choices made in a survey setting (i.e., choice experiments) are widely used to estimate the importance of product attributes in order to make product design and marketing mix decisions. Choice experiments allow the researcher to estimate preferences for product features that do not yet exist in the market. However, parameters estimated from experimental data often show marked inconsistencies with those inferred from the market, reducing their usefulness in forecasting and decision making. We propose an approach for combining choice-based conjoint data with individual-level purchase data to produce estimates that are more consistent with the market. Unlike prior approaches for calibrating conjoint models so that they correctly predict aggregate market shares for a "baseline" market, the proposed approach is designed to produce parameters that are more consistent with those that can be inferred from individual-level market data.

The proposed method relies on a new general framework for combining two or more sources of individual-level choice data to estimate a hierarchical discrete choice model. Past approaches to combining choice data assume that the population mean for the parameters is the same across both data sets and require that data sets are sampled from the same population. In contrast, we incorporate in the model individual characteristic variables, and assert only that the mapping between individuals' characteristics and their preferences is the same across the data sets. This allows the model to be applied even if the sample of individuals observed in each data set is not representative of the population as a whole, so long as appropriate product-use variables are collected that can explain the systematic deviations between them. The framework also explicitly incorporates a model for the individual characteristics, which allows us to use Bayesian missing-data techniques to handle the situation where each data set contains different demographic variables. This makes the method useful in practice for a wide range of existing market and conjoint data sets. We apply the method to a set of conjoint and market data for minivan choice and find that the proposed method predicts holdout market choices better than a model estimated from conjoint data alone or a model that does not include demographic variables.

**Keywords**
discrete choice modeling, conjoint analysis, choice experiments, data enrichment, hierarchical models, missing-data models, Bayesian estimation

**Disciplines**
Behavioral Economics | Business | Business Analytics | Marketing | Statistics and Probability

# Reality Check:
# Combining Survey and Market Data to Estimate Choice Models

Eleanor McDonnell Feit
Stephen M. Ross School of Business
University of Michigan
Ann Arbor, MI 48109
efeit@umich.edu

Mark A. Beltramo
Vehicle Development Research Lab
General Motors, Warren, MI 48090
mark.beltramo@gm.com

Fred M. Feinberg
Stephen M. Ross School of Business
University of Michigan
Ann Arbor, MI 48109
feinf@umich.edu

March 11, 2008

This material is based on the first author's dissertation and was supported in part by the National Science Foundation and General Motors under Grant No. DMI-0541610. Any opinions, findings and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the GM or the NSF. The authors thank Michel Wedel and Peter Lenk for their helpful suggestions.

**Abstract**

Discrete choice models estimated using hypothetical choices made in a survey setting (e.g., choice-based conjoint) are widely used to forecast the effects of product design and marketing mix decisions. Survey methods allow the researcher to estimate preferences for product features that do not yet exist in the market. However, parameters estimated from survey data often show marked inconsistencies with marginal effects inferred from the market, reducing their usefulness in forecasting and decision making. Several methods for adjusting survey-based choice models so that they more accurately predict market share have been suggested, but existing calibration methods are ad hoc and may change parameter values in ways that render them less consistent with other key empirical features of the data. We propose a new approach that produces more market-consistent parameter estimates by combining individual-level purchase data from the market with survey choice data in the formal estimation process.

The proposed method relies on a new general framework for combining two or more sources of choice data to estimate a hierarchical discrete choice model. Past approaches to combining choice data assume that the population mean for the parameters is the same across both data sets and require that data sets are sampled from the same population. In contrast, we incorporate individual demographic and product-use variables into the model and assert only that the mapping between individuals' demographics and their preferences is the same across the data sets. This allows the model to accommodate differences in choice behavior across the data sets driven by differences in observed demographics. The framework also explicitly incorporates a model for the individual demographics which allows us to use Bayesian missing data techniques to handle the situation where each data set contains different demographic variables. This makes the method useful in practice for a wide range of existing market and survey data sets. We apply the method to a set of conjoint and market data for minivan choice and find that the proposed method predicts holdout market choices better than a model estimated from conjoint data alone or a model that does not include demographic variables.

*Key words:* discrete choice modeling, conjoint analysis, choice experiments, data enrichment, hierarchical models, missing data methods, Bayesian estimation

For many companies, decisions made today about which products should be developed will drive profitability for years or even decades to come (Krishnan and Ulrich 2001). For example, a typical automotive product development program will invest hundreds of millions of dollars in design and tooling. In this capital intensive, mature market, the failure to meet sales targets by just a few percent can result in an unprofitable program (Urban, Hauser and Roberts 1990).

A rich array of methods has been devised to guide managers in their product design decisions. Such methods typically attempt to measure consumer preferences for various product attributes and use those measurements to make predictions about future individual-level purchase behavior in the market. The most successful and widely-applied among these is conjoint analysis (Green and Rao 1971), a set of experimental techniques that present consumers with various combinations of product attributes and statistically estimate consumer preferences for various levels of the attributes. These survey-based methods have proven their worth in a remarkable variety of contexts (for examples see Green, Krieger and Wind 2001), yet they are not without their drawbacks. Most notably, the attribute preferences estimated from conjoint tasks are sometimes inconsistent with preferences inferred from market data, an indication that respondents do not make hypothetical survey choices exactly as they make purchase decisions (Brownstone, Bunch and Train 2000, Blamey and Bennett 2001). One naïve response to this failing is to abandon conjoint and estimate models exclusively from market data. Unfortunately in practice it is difficult to estimate preferences for attributes directly from market data, primarily because there is insufficient variation in the products offered in the market, and the resulting parameter estimates sometimes lack face validity (c.f. Brownstone, Bunch and Train 2000). For example, is impossible to use market data to estimate preferences for new attributes that are not yet available in the market, which limits the value of market-data models for informing product design decisions. In this paper, we develop a general method for combining different sources of choice data and apply it to the specific problem of combining conjoint and market data. The resulting parameter estimates are informed by both the conjoint task and what is observed in the market and can be thought of as conjoint estimates with a "reality check".

**A new approach to combining choice data.** Methods for combining sources of preference data to estimate a *homogeneous* discrete choice model have been demonstrated in a number of applications in transportation research and environmental economics (see Ben-Akiva, Bradley, Morikawa et al. 1994 and Louviere, Meyer, et al. 1999 for reviews). The key modeling insight in past work is that combining two sets of choice data requires a scale parameter to accommodate differences in error scaling between them (Ben Akiva and Morikawa 1990, Swait and Louviere 1993). While there are many demonstrated benefits to combining sources of preference data, there remain a number of unresolved modeling issues concerning how to relate the two data sets together in the presence of consumer *heterogeneity (*Swait and Andrews 2003). In past work, researchers estimating heterogeneous choice models from multiple data sources required that each individual decision maker be observed making choices in both settings. With such data, they could impose the constraint that each individual maintained his or her preferences across the two choice contexts (Brownstone, Bunch and Train 2000, Bhat and Castelar 2002). However, in GM's experience, collecting such matched data would require effort in planning and recruitment that is impractical in commercial market research. Companies like GM that regularly conduct conjoint studies typically have access to individual-level market data that could readily be used to estimate joint models, but they seldom have this data for the *same individuals* that have completed the conjoint task. As we will discuss, the available data sources often have strengths that are complementary, and our goal is to build a flexible modeling framework that can be readily used with existing data collection methods and existing data.

A key element of our approach is a hierarchical choice model in which an individual's preferences depend on his or her personal characteristics (Allenby and Ginter 1995, Lenk et al. 1996). Incorporating individual characteristics in the joint modeling framework confers several advantages. Most importantly, if consumers taking part in a conjoint study differ systematically in a relevant way from those in the marketplace, the analyst should not impose the restriction that the distributions of preferences in these two groups are identical; doing so is an overt misspecification. So, instead of constraining the expected value of *preferences themselves* to be the same across the two data sets (as in Swait and Andrews 2003), we posit that the *relationship between individual-level characteristics and preferences* holds at the popu-

lation level, and so is the same across data sets drawn from that population. Individual-level characteristics can include not only relevant socioeconomic variables but also information from the consumer about what needs he of she desires the product to fulfill. We refer to the latter data collectively as *product-use variables.* When available, product-use variables are typically much more informative about preferences than are commonly available demographic data (Fennell et al. 2003, De Bruyn 2007).

Because we assume that the underlying relationships between the product-use variables and preferences are the same across data sets, the model structure can accommodate systematic differences in choice behavior between individuals observed in the market and those observed in the survey that can be related to the observed product-use variables. Thus, the approach can be applied even if the sample of individuals observed in each data set is not representative of the population as a whole, so long as appropriate product-use variables are collected that can explain the systematic deviations. Past approaches to combining choice data are restricted to data sets that are random samples from the same population (Swait and Andrews 2003), which is difficult to achieve in practice. For example market research conducted by GM often samples a group that has a somewhat different distribution of demographics than the population as a whole (e.g. automotive market research respondents are typically older than the general population of automotive buyers).

One potential disadvantage of incorporating individual characteristics using standard hierarchical choice models is that it requires that these variables are collected for all the decision makers observed in each choice data set. To maintain the assumption that the relationship between individual characteristics and attribute preferences is the same in each data set, it is critical that the same set of individual characteristics is accounted for in the regression of individual preferences on individual characteristics in both data sets. If the individual characteristics are correlated with one another, then omitting one of those individual characteristics may produce a bias in the remaining coefficients. Thus if the individual characteristic is omitted from one data set and not the other, the equality restriction on the parameters cannot be maintained. This is referred to as the omitted variables bias in the literature on meta-analysis of regression studies, c.f. Dominici et al. (1997). We overcome this disadvantage by incorporating a likelihood-based

approach to missing characteristics (Little and Rubin 2002), which allows us to include any individual characteristic observed in at least one of the data sets. Bayesian estimation proceeds in a natural fashion using data augmentation for the missing characteristics.

**Benefits of combining market and survey data.** The method we develop can be used to combine *any* available sources of choice data, such as choices observed at different retailers or in different research clinics, but there are particular advantages to combining survey choices with market data. The typical weaknesses in these two data sources are complementary, and combining them to estimate a single discrete choice model mitigates these weaknesses (c.f. Louviere, Hensher and Swait 2000, chapter 8).

**Table 1. Relative strengths and weaknesses of survey and market data.**

| Market Data | Survey Data (e.g. Conjoint) |
|---|---|
| ▪ Face validity<br>But,<br>▪ Inaccurate data on attribute values<br>▪ Missing some attributes of interest<br>▪ Colinearity between attributes<br>▪ Limited information on heterogeneity<br>▪ Limited information on consideration set<br>▪ Sample selection problems | ▪ Experimental design<br>▪ Complete record of choice<br>But,<br>▪ People may not choose in the same way they would in the market<br>▪ Information is not presented in the same way that it is in the market |

The crucial weakness of choice-based conjoint data (see Table 1) is that the hypothetical choices respondents make in the survey may not accurately reflect their behavior in the market. The conjoint task may provide the consumer with influential information that is not available in the normal shopping process. Worse, respondents may take cues from the conjoint design itself, placing undue emphasis on attributes they suspect are important to the researcher—e.g., the number-of-levels effect (Verlegh, Shifferstein and Wittink 2002). Consumers may also honestly believe that they place strong emphasis on socially-valued attributes (e.g., sustainability or 'green' technology) and make conjoint choices reflecting such values, yet ignore those attributes when making a purchase in the market (Blamey and Bennett 2001).

Market data obviously provides excellent face validity, but is plagued by problems that make statistical estimation using market data difficult. In a typical market many product attributes are correlated across product offerings, due to manufacturers' common objectives; such attribute colinearity makes it

difficult to parse out which attributes actually drive purchase behavior (Brownstone, Bunch and Train 2000). For example, if *every* minivan on the market that includes a navigation system also includes a rear-view camera, it is impossible in principle to disentangle these attributes' distinct effects based on purchase observations alone. For complex durables the situation is often exacerbated by physical design constraints. Because it is physically difficult to design a minivan that is both roomy and small, roominess and size are negatively correlated across minivans in the market. Most discrete choice models estimated from market data therefore restrict parameter estimation to marketing mix attributes (e.g., price, promotion, etc.) that tend to vary more independently than product design attributes. Market data is seldom used to estimate the importance of product design attributes (see Fader and Hardie 1996 for an exception). Furthermore, companies designing new products often wish to assess preferences for product attributes that are not yet available in the market, an impossibility using market data alone.

Combining data from a well-designed conjoint task with market data can improve the conditioning of the design matrix, providing the variation necessary to estimate main effects for all attributes of interest and resolving colinearity among attributes. Parameters that cannot be estimated from the market data, such as preferences for attributes that are not offered in the market, will be informed only by the conjoint data. When we combine conjoint and market data to estimate a joint model, the resulting parameters will be consistent with the conjoint choices unless there is sufficient evidence in the market data to make an adjustment. Other approaches calibrate conjoint models to the market by making *post hoc* adjustments to the estimated brand parameters so that the predicted shares closely match aggregate shares from the market (Orme and Johnson 2006, Gilbride, Lenk and Brazell 2006). Although these approaches only require aggregate share data, incorporating individual-level market data directly in the estimation is preferred, as it will not force changes to the conjoint parameter estimates that are not justified by the information contained in the market data. *Post hoc* calibration methods raise the possibility of 'overfitting' the aggregate shares and thereby performing poorly when predicting future shares.

There are additional benefits to combining conjoint and market data in the context of *heterogeneous* models. In market data for many product categories, we observe only one, or at most a few, choices

for each individual (Urban, Hauser and Roberts 1990). Without a sufficient number of choice observations per individual, it is difficult to estimate the amount of unexplained heterogeneity in preferences, even if the parametric models employed by the analyst are formally identified (Rossi, Allenby, and McCulloch 2005). By contrast, in a conjoint task it is relatively easy to collect multiple hypothetical choices for each respondent, and survey designs well-suited to estimating heterogeneous models can be readily developed (Sandor and Wedel 2005). When the data sets are combined, the conjoint data can serve to identify the distribution of heterogeneity, while still leveraging the observed preferences in the market data.

In the next section, we develop the model formally and explore its parameter recovery properties. Then we present an application of the model to survey and market choice data for the US minivan market collected by General Motors. In the final section, we summarize our conclusions and discuss future research directions.

**MODEL DEVELOPMENT**

Our model leverages the hierarchical discrete choice framework, where the part-worths of attributes are specified as a function of individual characteristics plus some error (Allenby and Gitner 1995, Lenk et al. 1996). We assume that each individual's choices are related to a vector of attribute preferences, $\beta_n$, $n \in \{1, \ldots, N\}$. These preferences follow a multivariate normal linear model, i.e.,

$$\beta_n = \beta_0 + \Delta z_n + \nu_n \qquad \nu_n \sim \mathrm{MVN}(\mathbf{0}, \Sigma_\nu) \tag{1}$$

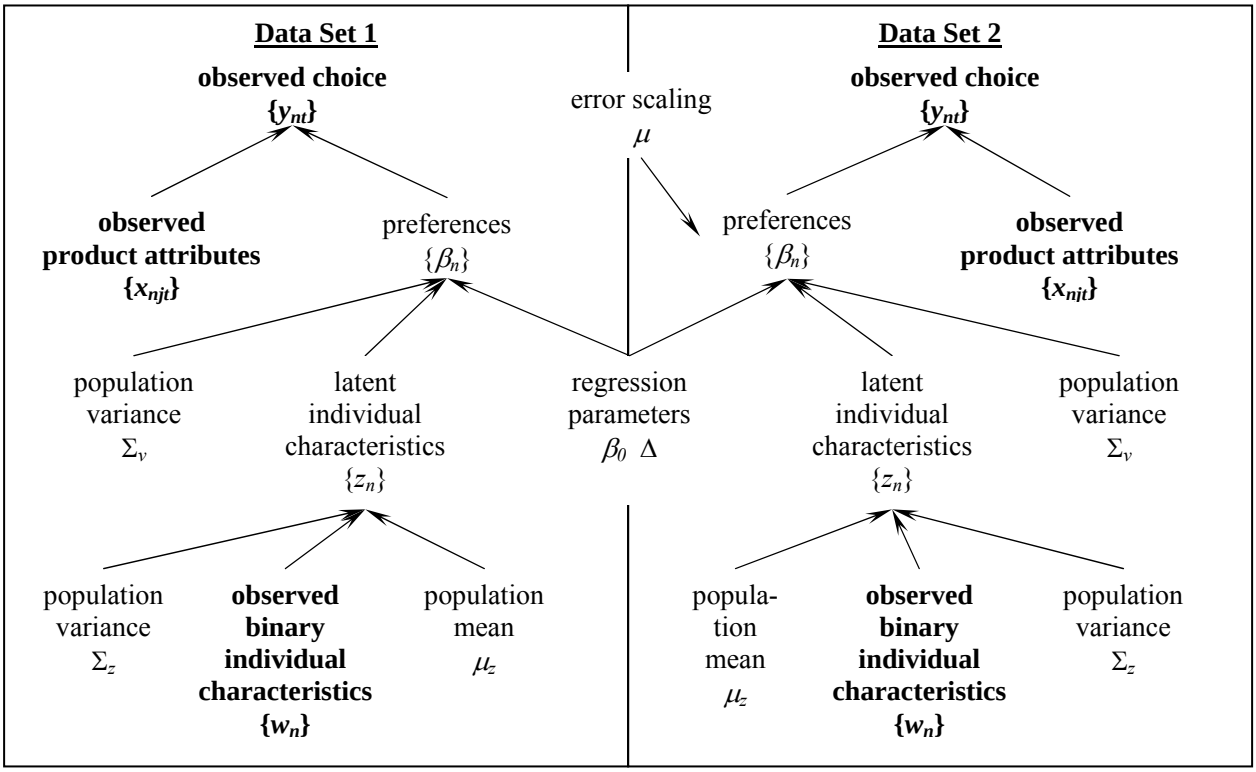
where $\beta_0$ is a vector of intercepts, $z_n$ is a vector of observed characteristics of the individual and $\Delta$ is an estimated matrix of regression parameters relating $z_n$ to $\beta_n$. The error term, $\nu_n$, is distributed multivariate normal with mean vector **0** and covariance matrix $\Sigma_\nu$. The data sets are related to one another by assuming that the parameters $\beta_0$ and $\Delta$ are common across data sets (see Figure 1). This implies that individuals with the same characteristics ($z_n$) will have the same expected value of $\beta_n$ and will therefore make

similar choices regardless of which data set they are observed in. This assumption is reasonable when the two data sets draw from the same population of decision makers.

By defining the relationship between the data sets through $\beta_0$ and $\Delta$, we also gain a great deal of flexibility. Past approaches to combining choice data do not include individual characteristics in this regression (i.e., they assume $\beta_n = \beta_0 + \nu_n$), which implies that the mean of $\beta_n$ is the same across the two choice contexts. This assumption requires that the two data sets both represent random samples from the target population, which can be achieved through careful sampling (Swait and Andrews 2003) or by collecting survey data for the same group of individuals as is observed in the market (Brownstone, Bunch and Train 2000, Bhat and Castellar 2002). In the proposed approach we can avoid assuming that all data sets are a random sample from the target population, if we have data to explain the relevant systematic differences between the two samples of decision makers. If there are differences in the distribution of $z_n$ between the data sets, then our model will predict that the distribution of $\beta_n$ and the resulting choices will be different across the two data sets. Because selection bias is prevalent in commercial marketing research, it seems prudent to accommodate any observed differences in the distribution of individual characteristics between the two samples. Of course, data with different empirical distributions of $z_n$ should only be combined in situations where the researcher is confident that the specification of the linear model is reasonable across the data sets. We do not recommend taking this approach to the extreme and combining, say, survey choices from 5 year-olds with market data from 55 year-olds in a situation where the relationship between age and preferences may be nonlinear (e.g., breakfast cereal choices.)

Figure 1 depicts how the preference parameters $\beta_0$ and $\Delta$ tie the two data sets together. In addition to this core multivariate regression, the model also includes a lower-level choice model relating product attributes to observed choices and a higher-level model describing the distribution of the individual characteristics. We describe these other components in more detail below.

**Figure 1. Proposed model structure.**




Conditional on a vector of preferences, $\beta_n$, we assume that the likelihood of observing a particular choice follows the standard random utility formulation. Specifically, we assume that on each choice occasion, $t \in \{1,...T_n\}$, individual $n$ will choose the alternative, $j \in \{1,...J_{nt}\}$, with greatest utility, $u_{njt}$, where

$$u_{njt} = \begin{cases} x_{njt}\beta_n + \varepsilon_{njt} & \text{if } n \in \text{data set 1} \\ \mu x_{njt}\beta_n + \varepsilon_{njt} & \text{if } n \in \text{data set 2} \end{cases} \quad \textbf{(2)}$$

and $x_{njt}$ is the row vector of attributes for alternative $j$ faced by decision maker $n$ on occasion $t$. Because we will be combining data sources where different sets of alternatives were included in the consideration set, we will assume that the error term, $\varepsilon_{njt}$, is distributed IID according to the standard Extreme Value distribution. The resulting model takes the familiar multinomial logit specification. In situations where the same set of alternatives appears in each choice task, a probit specification could be used to capture the error covariance between alternatives.

The Swait-Louviere scaling parameter, $\mu$ in equation 2, accounts for the possibility that the scale of the unexplained variation in $u_{njt}$ is different across the data sets (Ben-Akiva and Morikawa 1990, Swait and Louviere 1993). This can arise for a number of reasons. Individuals' choice consistency is known to vary across choice contexts (Bradley and Daly 1994): for instance, consumers may make more consistent choices when making real purchase decisions than when making hypothetical survey decisions. The scale parameter can also be used to accommodate situations where the set of observed attributes differs across choice sets (Swait and Louviere 1993). Thus, the scale parameters make it possible to combine data sources that have different, but overlapping, sets of product attributes, so long as the missing attributes are not correlated with observed attributes. Equation 2 describes the approach for combining two data sets; extensions to three or more data sets would incorporate an additional scale parameter for each additional data set and are straightforward.

It is crucial to note that, when this model is estimated, the resulting parameter estimates are not simply an 'average' (even suitably weighted) of those arising from each data set individually. For instance, when we combine market and conjoint data, there are often attributes that do not vary sufficiently in the market data. A near-orthogonal conjoint task can improve the conditioning of the $x_{njt}$ data; so, the resulting parameter estimates for those attributes will be based primarily on the conjoint data, where the data contains more information. The inclusion of conjoint data can even correct parameter estimates with incorrect signs caused by attribute colinearity in the market data (Brownstone, Bunch and Train 2000).

**Missing individual characteristics.** While the model described by equations 1 and 2 allows for a great deal of flexibility by incorporating individual characteristics, it can not be applied as-is when individual characteristics are correlated and there are missing individual characteristics in one data set. This limitation poses a serious challenge. Even within a company with a systematic marketing research program, like GM's, it is very seldom the case that precisely the same set of individual characteristics is available in both data sets. This is especially so when these individual characteristics are product-use questions (which are most likely to be informative about attribute preferences). The need for consistency

in ongoing market research programs (such as the survey of recent buyers that we use in our application) often limits opportunities for including new questions.

Because the method would be far less widely applicable if it required the same set of individual characteristics in both data sets, it is critical to overcome this apparent limitation. But this is made difficult by the intrinsically correlated nature of the individual characteristics themselves. When two correlated regressors affect a dependent variable in a linear regression, like that in equation 1, the omission of one regressor may produce biased estimates of the coefficient on the other regressor. Because of this omitted variables bias, it is difficult to combine regression data across multiple data sources when there are different sets of characteristics available in each data set (c.f. Dominici et al. 1997). If the regressors are correlated and each data set includes a different subset of the regressors, then regression coefficients for separately estimated models will both be biased in possibly different ways. This means the analyst can not assert that the regression coefficients are the same for those variables that are common across the two data sets. And if we cannot assert that the regression coefficients – $\beta_0$ and $\Delta$, in equation 1 – are the same across both data sets, then we no longer have a way to relate the two data sets together. Thus the model in equations 1 and 2 can not be applied in its stated form if different individual characteristics are available in each data set.

To address this problem, we adopt a likelihood-based approach to missing data. Unlike imputation approaches to missing data, likelihood-based approaches simply define the likelihood of the observed data as the marginal of the complete data likelihood, integrating over the distribution of the missing data. This marginal likelihood can be maximized or used in Bayesian inference. Specifically, if $[Y \mid \theta]$ is the likelihood of the complete data, $Y$, given some parameters $\theta$, and the data are missing at random (MAR), then the likelihood of an observed subset of the complete data, $Y_{obs} \in Y$, is the integral of the complete data likelihood over the missing data, $Y_{mis}$, i.e.,

$$[Y_{obs} \mid \theta] = \int_{Y_{mis}} [Y_{obs}, Y_{mis} \mid \theta] dY_{mis} \qquad \textbf{(3)}$$

This approach assumes that the process that caused the data to be missing is ignorable, which in turn requires that the characteristics are missing at random (MAR); that is, the probability that a covariate is missing does not depend on the realized value for that covariate, but may depend on other observed data. This is a reasonable assumption in the case where the covariate data simply was not collected in one of the data sources. It may not be appropriate in situations where particular respondents choose not to complete a particular survey question (e.g., if high income respondents are less likely to report their income). Importantly, the missing characteristics need not be missing completely at random (MCAR) for the approach to apply, and so the value of the missing characteristics *can* depend on the values of other characteristics or on the observed choices. (See Little and Rubin 2002 for a complete discussion.)

In the case of missing regressors, defining the complete data likelihood requires supplementing the usual likelihood for the dependent variables with a model for the regressors (Little and Rubin 2002, Dominici et al. 1997). The particular form of this model will depend on the nature of the regressors. For now, we will simply allow $[z_n \mid \varphi]$ to denote the likelihood of observing a covariate vector $z_n$ dependent on some parameters $\varphi$.

Given an expression for the likelihood of the individual characteristics and the assumption that they are each MAR, we can write the joint likelihood of the observed choices and the characteristics (assuming everything was observed). To simplify notation, let $y \equiv \{y_{11},...,y_{nt},...y_{NT_N}\}$ be the set of all observed choices, let $Z \equiv \{z_1,...,z_N\}$ be the (complete) set of individual characteristics, let $X_{nt} \equiv \{x_{n1t},...,x_{njt},...,x_{nJ_{nt}t}\}$ be the attribute data for a particular choice observation and let $X \equiv \{X_{11},...,X_{nt},...,X_{NT_N}\}$ be the set of all attribute data. The joint likelihood of $y$ and $Z$ conditional on $X$ and the parameters is

$$[y,Z \mid X,\mu,\beta_0,\Delta,\Sigma_\nu,\varphi] = \prod_n \left( \left( \int_{\beta_n} \left( \prod_{t\in\{1,...,T_n\}} [y_{nt} \mid X_{nt},\beta_n,\mu] \right) [\beta_n \mid z_n,\beta_0,\Delta,\Sigma_\nu] d\beta_n \right) [z_n \mid \varphi] \right) \qquad \textbf{(4)}$$

where the probabilities $[y_{nt} \mid X_{nt}, \beta_n, \mu]$ and the densities $[\beta_n \mid z_n, \beta_0, \Delta, \Sigma_v]$ are defined by equations 1 and 2.

Specifying the joint likelihood of *y* and *Z* as in equation 4 is consistent with the standard regression, which only specifies the likelihood of *y* conditional on *Z*. To see this, notice that the likelihood of $[z_n \mid \varphi]$ can be factored out of the joint likelihood in equation 4, and thus when *Z* is observed, the parameters that maximize the joint likelihood are the same as those that maximize the usual conditional likelihood.

When there are missing characteristics, we marginalize the complete data likelihood to obtain the likelihood of the observed data. Let $z_n^{mis}$ be the subset of $z_n$ that is missing and let $Z^{obs}$ be the subset of *Z* that is observed. The joint likelihood of observing a set of individual characteristics and a set of choices is

$$[y, Z^{obs} \mid \mu, \beta_0, \Delta, \Sigma_v, \varphi, X] = \prod_n \left( \int_{z_n^{mis}} \left( \int_{\beta_n} \left( \prod_{t \in \{1,\ldots,T_n\}} [y_{nt} \mid X_{nt}, \beta_n, \mu] \right) [\beta_n \mid z_n, \beta_0, \Delta, \Sigma_v] d\beta_n \right) [z_n \mid \varphi] dz_n^{mis} \right) \quad \textbf{(5)}$$

The integral over $z_n^{mis}$ makes it impossible to factor out the terms involving $[z_n \mid \varphi]$, so when there are missing characteristics, we must use this joint likelihood rather than the usual conditional likelihood. Once the model $[z_n \mid \varphi]$ is specified, the likelihood in equation 5 can be used in maximum likelihood or Bayesian estimation.

The model for $z_n$ can be any model that appropriately captures the relationships among the characteristics. If the $z_n$ are continuous with full support, they can be modeled via a multivariate normal model (Dominici et al. 1997). However, covariate data used in marketing is often discrete or measured using a discrete scale (e.g., employment status, income ranges). In our particular case study, the covariate data used was binary, so we illustrate the approach for binary data only; extensions to other common survey data types are analogous and straightforward. To model the vector of correlated binary characteristics, we

use a multivariate binary probit model (Chib and Greenberg 1998)[1]. We assume that the vector of zeros and ones that are observed, $w_n$, arises from an underlying multivariate normal vector, $z_n$ , as follows,

$$w_{nl} = \begin{cases} 1 & \text{if } z_{nl} > 0 \\ 0 & \text{if } z_{nl} \le 0 \end{cases} \quad \text{where, } z_n \sim \text{MVN}(\mu_z, \Sigma_z) \tag{6}$$

where $l$ indexes the elements of $z_n$ and $w_n$. The covariance matrix, $\Sigma_z$, is restricted so that the variance of each element of $z_n$ is one. The parameters of this model, $\mu_z$ and $\Sigma_z$, can be estimated separately for each data set (i.e. conjoint versus market). We model an individual's preference vector, $\beta_n$, as a function of the latent continuous vector, i.e., $\beta_n = \Delta z_n + \nu_n$ . This structure allows preferences to vary continuously as a function of the underlying constructs that gave rise to their binary responses, and also preserves conjugacy in the estimation algorithm.

**Estimation.** Our approach to estimation is Bayesian, using diffuse but proper priors on all parameters (c.f. Rossi, Allenby and McCulloch 2005). The integrals over $\beta_n$ and $z_n^{mis}$ in equation 5 are handled using data augmentation (Tanner and Wong 1987). The resulting Gibbs sampler draws sequentially from the posterior of the parameters $\beta_0$, $\Delta$, $\Sigma_\nu$, $\mu$, $\mu_z$, and $\Sigma_z$, and the unobserved latent variables $\beta_n$ and $z_n$ . The parameters of the multivariate probit model, $\mu_z$ and $\Sigma_z$, are sampled over the unidentified space and posterior distributions for the identified parameters are obtained by marginalizing over the posterior draws (McCulloch and Rossi 1994). The full conditional densities of all parameters are standard distributions, with the exception of $\beta_n$ and $\mu$, which were drawn using Metropolis-Hastings steps. Because $\beta_n$ is potentially a long vector, we used a normal random-walk proposal with an adaptive covariance matrix based on the covariance of all previous draws for individual $n$. This proposal density has been shown to maintain the convergence properties of the MCMC chain (Haario, Saksman and Tamminen 2001). The algorithm is described in detail in the e-companion.

[1] Alternatively, Ibrahim, Lipsitz and Chen (1999) propose to model a vector of missing discrete regressors as a series of related univariate generalized linear models.

**Insight into how the data informs the posterior for individual parameters.** An important characteristic of the model is that the posterior distribution of $\beta_n$ implied by the likelihood defined in equation (5) depends on both the observed choices, $\{y_{nt}\}$, and the individual characteristics, $z_n$. Incorporating characteristics informative to $\beta_n$ is critical to obtaining accurate individual-level parameters when only one or a few choices are observed for each respondent. For instance, knowing whether a household intends to use a minivan to transport children (in addition to which minivan model they chose) gives information about the household's preference for features like integrated DVD players. In data sets where a large number of choices are observed for each individual, such as conjoint data or long panel data, individual-level parameters can usually be well-recovered based only on the likelihood of the observed choices. But in market data for durables, where there is just one observed choice, incorporating individual characteristics that are related to choice behavior can significantly improve individual-level parameter recovery.

To demonstrate the value of individual characteristics, we generated a synthetic data set consisting of 10 choice observations for each of 100 'conjoint' respondents, and 1 choice observation for each of 1000 'market' respondents according to the complete-data model defined by equations 1-4. (Complete details of how the data was generated are included in the e-companion.) The data contained 5 individual characteristics, (i.e., $z_n$ was of length 5). Simulating the situation where some individual characteristics are not observed in one of the data sets, we assumed that the first characteristic was never observed for the conjoint respondents. We estimated the model using none, two or five individual-characteristics. (For example, in the case where two individual characteristics are used, the regression equation includes characteristics 1 and 2 and we observe characteristics 1 and 2 for the market individuals and just characteristic 2 for the conjoint individuals.)

Table 2 shows the improvement in individual parameter recovery that is gained by using individual characteristics. For each data set, we compute the mean squared error between individuals' true parameter value and their posterior mean for that parameter. We also compute the average standard error

around each individual's posterior mean, indicating how diffuse the individual posteriors are. Table 2 shows that as the number of individual characteristics included in the estimation is increased, the posterior means of $\beta_n$ for the market respondents approach their true values, even though we only observed one choice for each market respondent. Thus if $\beta_n$ is poorly identified by the observed choices, including informative characteristics improves posterior inference for $\beta_n$.

**Table 2. Recovery of individual-level betas for conjoint and market respondents.**

| Number of binary individual characteristics observed in market data | Recovery of $\beta_{n1}$ | | | | Recovery of $\beta_{n2}$ | | | |
|---|---|---|---|---|---|---|---|---|
| | Conjoint | | Market | | Conjoint | | Market | |
| | MSE | Avg. SE | MSE | Avg. SE | MSE | Avg. SE | MSE | Avg. SE |
| 0 | 2.37 | 1.19 | 4.04 | 1.75 | 1.53 | 1.05 | 3.33 | 1.64 |
| 2 | 2.09 | 1.30 | 3.34 | 1.75 | 1.40 | 1.05 | 1.98 | 1.42 |
| 5 | 1.71 | 1.21 | 2.15 | 1.47 | 1.37 | 1.04 | 1.92 | 1.42 |

More importantly, we also find that *increasing the number of observed choices improves inference about any missing individual characteristics*. Because we use a likelihood-based approach to the missing individual characteristics, our inference about a particular individual's latent characteristics ($z_n$) is informed both by what we know about other individuals' characteristics *and* the choices we have observed for that individual. In fact, the likelihood of $z_n$, conditional on the observed data and the other parameters, depends on both the model for the characteristics *and* on the choice parameters, $\beta_n$, as follows:

$$[z_n \mid w_n, \mu_z, \Sigma_z, \beta_0, \Delta, \Sigma_v] \propto [\beta_n \mid z_n, \beta_0, \Delta, \Sigma_v][z_n \mid w_n, \mu_z, \Sigma_z] \quad \textbf{(7)}$$

This stands in contrast to other approaches to missing data, such as multiple imputation and hot-deck, where missing characteristics would be imputed based only on information about other respondents' characteristics, ignoring the observed choices of the respondent in question.

To demonstrate the importance of observed choices in imputing the latent $z_n$ and missing elements of $w_n$, we conducted a second synthetic data study. Using the same 100 conjoint respondents and 1000 market respondents, we assumed that two binary individual characteristics were observed for each market respondent ($w_{n1}$ and $w_{n2}$) and that just one characteristic ($w_{n2}$) was observed for the conjoint respondents. We estimated the model four times, with three, ten, fifty and one hundred observed choices for

each conjoint individual. Table 3 shows that as the number of observed choices increases, recovery of the latent continuous variable $z_{n1}$ improves for the conjoint respondents. As the number of choice observations increases, the posterior distributions for $z_{n1}$ becomes less diffuse with means closer to the true values. Inference for the second covariate, $z_{n2}$, for which we observe $w_{n2}$ for the conjoint respondents, also improves slightly.

**Table 3. Recovery of individual characteristics for conjoint and market respondents.**

| Number of choices observed for conjoint respondents | Recovery of $z_{n1}$ | | | | Recovery of $z_{n2}$ | | | |
|---|---|---|---|---|---|---|---|---|
| | Conjoint | | Market | | Conjoint | | Market | |
| | MSE | Avg. SE | MSE | Avg. SE | MSE | Avg. SE | MSE | Avg. SE |
| 3 | 0.87 | 1.08 | 0.36 | 0.75 | 0.35 | 0.56 | 0.32 | 0.59 |
| 10 | 0.79 | 1.00 | 0.36 | 0.70 | 0.31 | 0.52 | 0.32 | 0.62 |
| 50 | 0.74 | 1.08 | 0.37 | 0.83 | 0.25 | 0.56 | 0.33 | 0.71 |
| 100 | 0.69 | 0.93 | 0.36 | 0.72 | 0.22 | 0.49 | 0.32 | 0.63 |

## APPLICATION: US MINIVAN MARKET

General Motors is among the many companies that regularly use choice models based exclusively on conjoint data to predict how new products will perform in the market. Because GM managers use these models to make critical product development decisions, they are keenly interested in improving the predictive accuracy of these models. Methods that can be applied to existing conjoint data, with minimal additional data collection, are extremely valuable to GM and other practitioners, as they can be applied in situations where a conjoint study has been fielded, and the resulting parameter estimates are found to lack face validity. Because our model can accommodate data where different groups of respondents are observed in the conjoint setting and in the market setting, we can readily augment an existing conjoint data set with purchase data from a different set of consumers.

In this section we describe how our method was used to adjust the parameters of a conjoint model for minivan purchase. Not surprisingly, the model estimated jointly from conjoint and market data fits the market data better than a model estimated from conjoint data alone and is therefore more useful when making predictions about the market. We also explore the value of including the product-use variables in the formulation by comparing our model to a model estimated without individual-level characteristics.

**Conjoint and market data.** The conjoint data for this application is a subset of data collected for a large conjoint study that was designed and fielded by GM in summer 2003. Our subset consists of 12 choice responses for each of 199 respondents who were selected based on their interest in purchasing a new minivan. In each choice task the respondents chose from among three alternatives with three attributes: price (levels: $20,000, $23,000, $26,000, $29,000, $32,000, $35,000), styling appeal (levels: very unappealing, unappealing, neutral, somewhat appealing, very appealing) and brand (14 levels which we label A-M at the request of GM). Respondents were randomly assigned to one of two fixed designs and made forced choices from among three alternatives. The choice questions were designed by GM's conjoint vendor using a proprietary method that allows for efficient estimation of a heterogeneous multinomial logit model. Although product-use variables were not systematically collected for each respondent in the conjoint study, the demographic profile did include one variable that is related to minivan product needs: number of children in the household.

To assemble market data that could be combined with the original conjoint study, we drew on an ongoing GM-proprietary survey of new vehicle buyers. This mail-out survey is sent quarterly to a sample of all new vehicle registrants. We selected from this survey all of the 7078 respondents who purchased a minivan during the 2004 model year (September 2003 – August 2004). For each respondent we observed one choice (the minivan purchase that qualified them for the survey) from among the 12 minivans that were on the market in 2004. The attribute data for the 12 minivans on the market were assembled from several sources. The average consumer price paid (negotiated price less consumer rebates) for each of the minivan models was estimated based on the price reported by other buyers in the same survey. Although the prices faced by a particular individual (who may have been a particularly good negotiator, or shopping for a minivan with many extra features) could be different than the average prices we use in estimation, we assume that the average prices reasonably reflect the *relative* prices faced by each respondent.[2] These

[2] Because vehicle prices in the US are privately negotiated between the buyer and the dealer, it is difficult for a manufacturer like GM to get transaction data for the particular individual that they have surveyed. Transaction data collected by third parties, such as J.D. Power PIN, does not contain the informative product-use variables that were included in the GM survey of recent buyers.

averages were computed by month to reflect seasonal price variation in the market data. We also assembled data from another GM source on the average consumer-rated styling appeal of each van (on the same scale as the conjoint study). In addition to the three variables included in the conjoint task, we also chose to include the attribute "date of last design refresh" in the market data. Newer designs tend to have better features and performance as well as higher prices. By controlling for age of the design, we hoped to get a more accurate estimate of the importance of price in the market data. The attribute data for the minivans on the market is summarized in Table 4. Brands D and M are excluded from the market data, since there were no minivans on the market from brands D and M. Brands A and B launched re-designed products in April 2004 so we have used different attribute data for the old and the new designs.

**Table 4. Attribute data for alternatives available in the market.**

| Brand | | N | K | J | B (old) | B (new) | H | A (old) | A (new) | E | L | I | C | F | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Styling Appeal | | 3.48 | 3.52 | 3.29 | 3.59 | 3.59 | 3.47 | 4.03 | 4.03 | 3.67 | 3.65 | 3.46 | 3.47 | 3.31 | 3.27 |
| Price ($K) | Oct 03 | 16.8 | 18.8 | 20.4 | 16.3 | NA | 22.8 | 20.7 | NA | 18.8 | 21.2 | 21.4 | 22.0 | 12.7 | 17.5 |
| | Nov 03 | 16.4 | 19.7 | 20.4 | 17.5 | NA | 22.3 | 18.2 | NA | 19.0 | 22.0 | 20.5 | 21.3 | 12.8 | 16.3 |
| | Dec 03 | 16.3 | 18.1 | 19.8 | 15.2 | NA | 20.8 | 17.8 | NA | 18.1 | 20.8 | 21.1 | 20.4 | 12.6 | 15.4 |
| | Jan 04 | 17.6 | 19.4 | 20.3 | 16.4 | NA | 23.2 | 16.9 | NA | 19.3 | 21.8 | 20.9 | 20.4 | 11.8 | 14.5 |
| | Feb 04 | 17.1 | 19.6 | 20.8 | 15.3 | NA | 22.9 | 16.6 | NA | 18.5 | 22.0 | 22.5 | 18.5 | 11.7 | 14.4 |
| | Mar 04 | 16.6 | 18.5 | 19.0 | 16.0 | NA | 19.4 | 20.1 | NA | 18.1 | 20.8 | 21.4 | 20.7 | 12.0 | 15.9 |
| | Apr 04 | 17.6 | 18.8 | 22.2 | NA | 16.6 | 21.5 | NA | 21.0 | 18.3 | 18.9 | 22.0 | 17.8 | 10.3 | 12.9 |
| | May 04 | 18.2 | 18.4 | 20.6 | NA | 16.3 | 19.6 | NA | 18.7 | 17.1 | 21.4 | 21.1 | 19.0 | 10.4 | 13.9 |
| | Jun 04 | 16.0 | 19.2 | 19.1 | NA | 17.7 | 24.4 | NA | 19.4 | 17.5 | 21.1 | 22.3 | 18.7 | 10.2 | 13.7 |
| | Jul 04 | 15.3 | 18.8 | 19.6 | NA | 16.3 | 19.6 | NA | 18.7 | 18.8 | 20.8 | 20.6 | 16.3 | 11.2 | 15.1 |
| | Aug 04 | 14.5 | 16.4 | 17.8 | NA | 15.4 | 17.8 | NA | 18.0 | 17.4 | 21.4 | 22.9 | 19.6 | 11.9 | 13.4 |
| | Sep 04 | 16.4 | 17.0 | 18.6 | NA | 14.6 | 15.7 | NA | 19.3 | 18.8 | 22.9 | 16.8 | 12.5 | 11.7 | 15.3 |
| Last Design Refresh | | 1997 | 1997 | 1997 | 2001 | 2005 | 2004 | 2001 | 2005 | 1999 | 2004 | 2004 | 2004 | 2002 | 2000 |

Because we have only one choice observation for each respondent in the market, it is helpful to incorporate individual characteristics for the buyers in the market data. To improve individual-level parameter recovery, such characteristics should be correlated with attribute preferences and observed choices. Although past research has found relatively little correlation between standard demographic variables and attribute preferences (Fennell et al. 2003), variables that capture information about intended product usage or product needs have been found to be highly correlated with product choices (De Bruyn et al. 2007). We were able to construct similarly informative individual characteristics using a section from the market survey where respondents could check off any of 78 potential "reasons for purchase", such as

"Luggage/cargo capacity" and "Family oriented". GM developed these questions over several years of fielding the survey; the reasons were designed to be an exhaustive set and GM had found that they were related to brand choice. GM grouped these 78 items into 25 blocks of similar reasons using a clustering approach that resulted in groups with high face validity (see Table 5). Using these blocks, we coded a binary variable for each respondent indicating whether the respondent had selected any item in the block. These 25 binary reasons-for-purchase variables entered the model as characteristics to the brand parameters. (We excluded from the data 208 buyers who did not check any of the 76 reasons, indicating that they failed to respond to that section of the survey.) In addition to the 25 reasons-for-purchase variables, we also included binary variables for whether the household had income less than $75K per year (roughly the median in this sample) and for whether or not the household had children as covariates to the price parameters.

**Table 5. Summary of individual characteristics.**

| Covariate | % of Respondents | | Covariate | % of Respondents | |
|---|---|---|---|---|---|
| | Field | Conjoint | | Field | Conjoint |
| Household Income <75K | 54.2% | - | Towing / Hauling | 6.2% | - |
| Household with Children | 48.8% | 48.4% | Accident Safety | 65.3% | - |
| Usability | 75.0% | - | Collision Avoidance | 35.3% | - |
| Dependability | 63.6% | - | Kid Features | 51.4% | - |
| Rugged / AWD / RWD | 12.0% | - | Exterior Styling | 61.0% | - |
| Dealer | 57.0% | - | Fun to Drive | 33.7% | - |
| Warranty | 44.8% | - | Country of Origin | 20.2% | - |
| Roominess | 78.5% | - | Practical | 51.8% | - |
| Cargo / Versatility | 52.2% | - | Environment | 15.4% | - |
| Fuel Economy / Value | 58.7% | - | Manufacturer Reputation | 58.9% | - |
| Incentives | 43.6% | - | Interior styling | 59.2% | - |
| Driving Performance | 59.8% | - | Willing to Negotiate | 30.6% | - |
| No Negotiation | 34.8% | - | Cargo Loading | 35.6% | - |
| Luxury | 25.2% | - | | | |

Because the reasons for purchase data and the income data were not collected for the conjoint respondents, the likelihood-based missing data approach was used to account for these missing variables. As there is but one individual characteristic common across the two data sets, this represents a fairly extreme instance of missing characteristics. Because there is no information available to estimate $\mu_z$ and $\Sigma_z$ separately for the conjoint data set, we took them to be common across the two data sets. Note, however, that we observe a relatively large number of well-designed choices for the conjoint respondents and based

on these choices, the posterior for $\beta_n$ is often quite tight. (The conjoint study was, after all, designed to infer $\beta_n$ from the choices.) When the posterior for $\beta_n$ is tight, it is possible that the posterior for $z_n^{mis}$ will also be quite tight and the distribution of $z_n$ for the conjoint respondents may be different from the distribution implied by $\mu_z$ and $\Sigma_z$.

In the interest of parsimony, we placed restrictions on which individual characteristics were included in the regression for each attribute preference. For instance, whether or not a respondent has children or high income is excluded from the model of a respondent's preferences for particular minivan brands; the effect on brand preference of the former is captured by the "Kid Features" and other reasons for purchase variables and the effect of the latter operates through its effect on price sensitivity. Also, we assumed that the reasons for purchase are not related to the respondent's price sensitivity.

The ongoing GM survey from which we collected the market data samples respondents on the basis of their chosen vehicle. The survey is mailed to a stratified sample of owners who have registered a new vehicle during the year, with the goal of receiving a fixed number of returns for each vehicle model. All returned surveys (typically around 20-25% of those mailed out) are included in the data set. To approximately adjust this choice-based sample to known market shares from national registration data, we adjust the likelihood of each individual's choice following Manski and Lerman (1977), as follows

$$[y_{nt} \mid X_{nt}, \beta_n, \mu] = \frac{\exp\left(x_{n,y_{nt},t}\beta_n + \log(s_{y_{nt},t})\right)}{\sum_j \exp\left(x_{njt}\beta_n + \log(s_{jt})\right)}, \qquad \textbf{(8)}$$

where $s_{jt}$ is the sales-to-sample ratio for alternative *j* at time *t.* (Manski and Lerman developed this correction for homogeneous choice models.) Sales-to-sample ratios were computed for each month based on the number of survey responses and national sales data.

The resulting market data consisted of 6870 respondents for whom we observed one purchase and the 27 individual characteristics. We divided this data set into 2356 randomly selected individuals to be used for estimation with the remainder reserved as a holdout sample.

**Model estimates.** The parameters of the model estimated from the minivan conjoint and market data are shown in Table 6. All reported estimates are based on 100,000 draws from each of two chains thinned to every 10th draw. Convergence was assessed by comparing the two chains and nearly all of the monitored parameters achieved Gelman-Rubin potential scale reduction factors below 1.1 (Brooks and Gelman 1998). Trace plots comparing the log-likelihood of the draws also indicated that the two chains had converged.

**Table 6. Estimated parameters for joint market/survey model.**

| | ***mu*** | ***Delta*** | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **0.55** | Other Attributes (X) | | | | Brands (X) | | | | | | | | | | | | |
| | | Styling Appeal | Price (linear) | Price (squared) | Design Age | A | B | C | D | E | F | G | H | I | J | K | L | M | ***mu.z*** |
| *Covariates (z)* | Intercept | **2.74** | **-1.43** | **-0.24** | **1.08** | **3.92** | **2.07** | 0.40 | **0.35** | -0.28 | **-6.17** | **-3.43** | -0.04 | -0.91 | **2.95** | 0.77 | 0.12 | **-1.69** | **NA** |
| | Household.Income.<75 | | **-1.00** | 0.25 | | | | | | | | | | | | | | | **0.13** |
| | Household.with.Children | | **-0.65** | 0.05 | | | | | | | | | | | | | | | -0.04 |
| | Reason.Usability | | | | | 0.38 | **1.02** | 0.06 | | **-1.26** | -0.01 | -0.59 | 0.02 | -0.58 | 0.69 | 0.39 | -0.88 | | **0.67** |
| | Reason.Dependability | | | | | **-0.93** | -0.27 | -0.05 | | **1.59** | 0.36 | 0.64 | -0.60 | 0.63 | -0.35 | **-1.18** | **1.48** | | **0.37** |
| | Reason.Rugged_AWDRWD | | | | | -0.33 | -0.58 | -0.25 | | **-0.87** | **-1.04** | -0.27 | -0.79 | **-1.05** | **1.33** | **1.57** | 0.25 | | **-1.17** |
| | Reason.Dealer | | | | | -0.12 | -0.05 | **1.11** | | **-0.81** | -0.09 | 0.02 | **1.48** | **-1.16** | 0.08 | 0.43 | **-1.45** | | **0.17** |
| | Reason.Warranty | | | | | 0.16 | **0.73** | **-0.92** | | -0.09 | **3.24** | **0.85** | **-1.02** | -0.25 | 0.08 | **-1.42** | **-0.85** | | **-0.13** |
| | Reason.Roominess | | | | | -0.38 | -0.59 | **-0.73** | | 0.28 | 0.16 | **0.85** | -0.28 | 0.82 | -0.70 | 0.39 | 0.08 | | **0.79** |
| | Reason.Cargo_Versatility | | | | | 0.30 | 0.12 | **0.73** | | 0.02 | **-1.34** | -0.30 | 0.48 | -0.16 | 0.23 | 0.10 | 0.08 | | **0.06** |
| | Reason.FuelEcon_Value | | | | | **-0.72** | -0.08 | **-1.07** | | 0.19 | 0.80 | **1.41** | **-0.86** | 0.14 | **-0.54** | 0.23 | 0.51 | | **0.22** |
| | Reason.Incentives | | | | | 0.49 | 0.47 | **1.57** | | **-2.50** | -0.05 | **-0.69** | **1.57** | **-2.68** | **1.70** | **1.81** | **-3.07** | | **-0.18** |
| | Reason.DrivePerform | | | | | 0.61 | -0.13 | 0.74 | | **-1.06** | -0.10 | -0.68 | 0.36 | **-1.28** | 0.70 | 0.76 | -0.65 | | **0.25** |
| | Reason.No_Negotiation | | | | | 0.07 | **-0.65** | -0.23 | | 0.43 | 0.75 | -0.21 | -0.11 | 0.40 | -0.37 | 0.18 | -0.10 | | **-0.39** |
| | Reason.Luxury | | | | | **0.74** | -0.45 | -0.18 | | -0.37 | -0.47 | -0.32 | **0.85** | **1.54** | -0.45 | **-1.28** | **1.00** | | **-0.66** |
| | Reason.Tow_Haul | | | | | 0.42 | 0.32 | -0.25 | | 0.08 | 0.81 | 0.18 | -0.34 | 0.32 | -0.52 | -0.64 | 0.07 | | **-1.52** |
| | Reason.Safety_Security | | | | | -0.62 | -0.54 | 0.80 | | 0.47 | 0.72 | 0.18 | 0.39 | 0.00 | **-1.03** | **-1.41** | **1.02** | | **0.40** |
| | Reason.AvoidCollision | | | | | **-1.01** | **-1.11** | 0.61 | | -0.08 | -0.64 | -0.67 | **1.49** | 0.74 | 0.49 | **1.07** | -0.04 | | **-0.37** |
| | Reason.KidFeatures | | | | | -0.14 | 0.21 | -0.34 | | **0.96** | -0.56 | 0.15 | **-1.20** | **0.92** | **-0.55** | -0.03 | **0.59** | | 0.03 |
| | Reason.ExteriorStyling | | | | | -0.12 | 0.08 | -0.11 | | -0.57 | -0.49 | **0.65** | -0.03 | 0.46 | 0.10 | 0.52 | **-0.71** | | **0.28** |
| | Reason.FuntoDrive | | | | | **-1.37** | 0.02 | **-1.17** | | 0.28 | 0.75 | **1.35** | -0.60 | **2.40** | -0.49 | -0.08 | -0.04 | | **-0.41** |
| | Reason.CountryofOrigin | | | | | 0.38 | **0.63** | 0.14 | | **-1.08** | **-1.74** | **-1.02** | 0.19 | **-0.85** | **1.25** | **1.44** | **-0.96** | | **-0.83** |
| | Reason.Practical | | | | | 0.32 | 0.30 | 0.11 | | -0.29 | 0.38 | -0.52 | -0.11 | **-0.89** | 0.20 | **0.64** | -0.48 | | 0.04 |
| | Reason.Environment | | | | | **0.91** | 0.39 | 0.34 | | 0.75 | -0.59 | -0.69 | 0.03 | -0.80 | -0.10 | -0.40 | 0.39 | | **-1.01** |
| | Reason.MfgReputation | | | | | 0.35 | 0.08 | -0.71 | | **2.23** | **-1.75** | -0.40 | -0.76 | 0.16 | 0.03 | **-1.09** | **1.96** | | **0.24** |
| | Reason.InteriorStyling | | | | | 0.23 | -0.48 | -0.20 | | 0.11 | **1.57** | 0.00 | 0.25 | -0.12 | -0.35 | -0.16 | 0.43 | | **0.23** |
| | Reason.WillingtoNegotiate | | | | | 0.31 | **0.54** | -0.43 | | 0.37 | -0.25 | 0.01 | -0.17 | 0.20 | -0.06 | 0.00 | -0.24 | | **-0.50** |
| | Reason.CargoLoading | | | | | 0.01 | 0.25 | 0.19 | | -0.01 | -0.51 | 0.30 | -0.23 | 0.31 | -0.22 | -0.53 | 0.58 | | **-0.37** |
| ***Sigma.nu*** | | **3.59** | **5.88** | **0.90** | **3.08** | **7.52** | **11.13** | **4.94** | **1.29** | **2.48** | **2.68** | **3.71** | **2.07** | **3.78** | **2.64** | **6.29** | **1.91** | **10.88** | |

The central panel in Table 6 shows the parameters in Δ describing the relationship between the individual characteristics and the choice parameters. Consistent with intuition, the intercept for the Styling Appeal parameter is positive and the intercept for the Price (linear) parameter is negative. Households with lower than median income and households with children have higher price sensitivity. Many of the other parameters in Δ are consistent with GM managers' intuition, for example, the estimate for the rela-

tionship between the "Warranty" reason for purchase and preference for brand "F" is high, indicating that respondents for whom warranty is important are more likely than others to choose brand F. This is consistent with brand F's industry-leading warranty program. We also find that buyers who indicated "Country of Origin" as a reason for purchase had significantly lower preferences for brands E, F, G, I and L, which were the only non-US brands in the sample.

The last row in Table 6 lists the estimated variances of the unexplained population heterogeneity for the attributes. There is more unexplained heterogeneity in preferences for styling and the linear term for price and less unexplained heterogeneity in preferences for age of vehicle design and the squared term for price. Heterogeneity in brand preferences varies widely depending on the brand. Some brands seem to be more universally liked or disliked while others appear to have more dispersion across individuals. The right-most column in Table 6 shows the estimated population means for the multivariate probit model that was used to estimate missing individual characteristics.

**The value of incorporating market data.** The key benefit of the proposed modeling framework is that it allows us to incorporate both market and conjoint data to estimate a choice model. By combining these sources of data, the resulting model still benefits from the well-conditioned attribute data in the conjoint study, yet should make more accurate predictions about choices in the market. To gain some insight into the effect of incorporating the market data, we compare the joint model to a model estimated from the conjoint data alone. We compare the ability of the joint model and the conjoint model to predict market data based on the posterior predictive likelihood of the estimation and holdout choices. We compute the log posterior predictive likelihood (*lppl*) of an observed choice by individual *n* on occasion *t* as

$$\log\left(\int_{\beta,z,\mu,\beta_0,\Delta,\Sigma_\eta,\mu_z,\Sigma_z} [y_{nt} \mid X_{nt},\beta,\mu][\beta \mid z,\beta_0,\Delta,\Sigma_v][z \mid w_n,\mu_z,\Sigma_z] \right.$$
$$\left. [\mu,\beta_0,\Delta,\Sigma_v,\mu_z,\Sigma_z \mid \text{data}]d\beta dz d\beta_0 d\Delta d\Sigma_\eta d\mu_z d\Sigma_z\right) \quad \textbf{(9)}$$

where $[\mu,\beta_0,\Delta,\mu_z,\Sigma_z \mid \text{data}]$ is the posterior distribution of the population parameters. (Note that *lppl* is proportional to the deviance averaged over the posterior distribution of the population parameters.) The average hit rate associated with this probability can be estimated by exp(*lppl*/*N*) where *N* is the number of

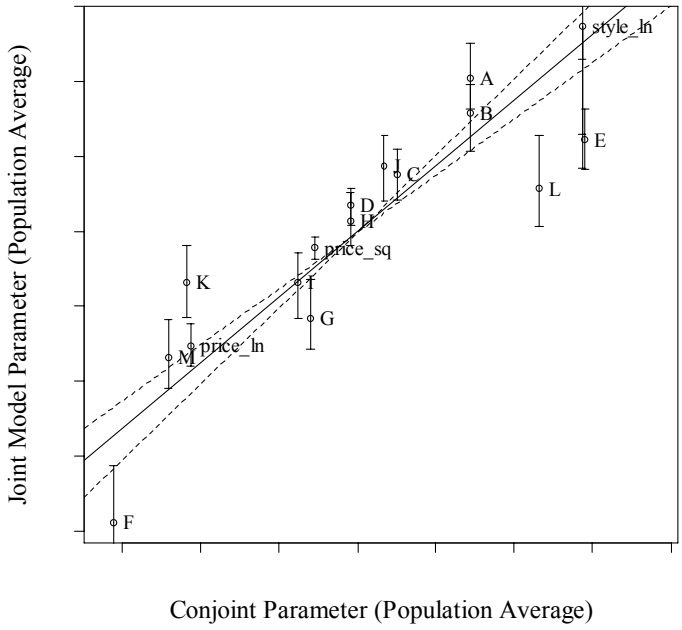
choice observations. Note that the conjoint and the joint models have essentially the same *structure* and differ primarily on what *data* is used in estimation. Thus it would be less appropriate to compare the models using Bayes Factors, which compare how well two models with different structure fit the same data set.

In this example, we find that the joint model does a substantially better job at predicting market choices than the conjoint model. The *lppl* of the market choices used in estimation is -4711 (average hit rate = 13.5%) for the joint model versus -6333 (average hit rate = 6.8%) for the conjoint model. In fact, the conjoint model does worse at predicting market data than a model that predicts according to the aggregate shares in the market data (*lppl*=-5830, average hit rate = 8.4%), an indication that the preferences expressed in the conjoint study were inconsistent with market shares. Clearly, incorporating market data in the estimation is essential to producing a model that will make accurate predictions about the market. We find a similar pattern of results in the *lppl* of the holdout market data (see Table 7.) Interestingly, we also find that the joint model only does slightly worse than the conjoint model at predicting conjoint choices (*lppl*=-2325 versus *lppl*=-2249.) Thus, the joint model can predict market data much better than the conjoint model, yet still makes reasonably good predictions for the conjoint data.

To understand why the conjoint model makes such poor predictions for the market data, we compare the parameters of both models. Figure 2 shows the differences between the population mean of $\beta_n$ estimated from the conjoint model versus the joint model. (When conjoint data are analyzed alone, the specifications of $\beta_0$, $\Delta$ and $\Sigma_\nu$ differ from that in the joint model due to the missing characteristics in the conjoint data. We therefore compare the distributions of individual $\beta_n$ implied by each model, but we do not compare $\beta_0$, $\Delta$ and $\Sigma_\nu$ directly. The conjoint-only parameter estimates for $\beta_0$, $\Delta$ and $\Sigma_\nu$ are included in the e-companion.) The population means of the elements of $\beta_n$ are plotted for the joint model versus the conjoint model and the error bars indicate the 5$^{th}$ and 95$^{th}$ percentile for the joint model estimates.

The solid diagonal line in Figure 2 has a slope equal to the estimated scale difference ($\mu$) between the conjoint and the market data and the dashed lines show the 5th and 95th percentiles of the posterior distribution of $\mu$. When a parameter falls below this range, it suggests that the attribute level is less preferred in the joint model relative to the conjoint model. Parameters above the range are associated with attribute levels that are more preferred in the joint model. Although the joint model can change the parameters for styling and price, substantial adjustments to these parameters do not seem to be supported by the data. The joint model does substantially adjust preference for many of the brands, which we attribute to differences between the general brand attitudes that consumers express in the survey setting versus how the brands are perceived when the customer is shopping and is actively engaged in collecting information about the brands and specific products. Important product planning decisions, such as which brands should offer minivans in the future, would be misinformed by predictions made using the model estimated from conjoint data alone.

**Figure 2. Comparison of estimated parameters for conjoint and joint models.**




A key benefit of using the market data directly in the estimation of the model is that the parameter estimates obtained will only be adjusted away from the conjoint model *if the information in the market*

*data justifies an adjustment*. This contrasts with calibrations that are done after estimation that require the researcher to specify which parameters should be adjusted and may erroneously adjust the wrong parameters in order to improve aggregate share predictions. In our approach it is the data that determines which parameters are adjusted; substantial differences between the conjoint and joint model suggest which attributes might require better stimuli (i.e. more similar to the market) in future conjoint studies.

**The value of incorporating individual characteristics.** To assess the benefits of incorporating individual characteristics, we also compare our model to one where all elements of $\Delta$ are fixed to zero, leaving only the intercepts. We will refer to this specification as the No-Individual-Characteristics (NIC) formulation. Similar to the model of Swait and Andrews (2003), this model forces the means of the part worths to be the same across the two data sets. We find overwhelming support for our model over the NIC model. The Newton-Raftery estimator (Newton and Raftery 1994) of the marginal likelihood of the NIC model is -4216 versus -3273 for the proposed model, indicating a log Bayes Factor of -943 in favor of our proposed formulation. (Because of the unbounded sampling variance of the Newton-Raftery estimator, we repeated this calculation individually for each chain. We found the estimated integrated log-likelihood to be similar across the chains with a difference of no more than 40 points, still clearly favoring our proposed formulation.) We also found that the *lppl* of the holdout data is significantly better for our formulation (*lppl*=-9689 for the joint model versus -11100 for NIC model). In other data sets where the differences in the distribution of individual characteristics between the market and the conjoint data are better observed, we would expect our formulation to be even more strongly favored.

**Table 7. Log Posterior Predictive Likelihoods and Average Hit Rates**

| Model | Fit to Estimation Data | | | | Fit to Holdout Data | |
|---|---|---|---|---|---|---|
| | Survey | | Market | | | |
| | *lppl* | average hit rate | *lppl* | average hit rate | *lppl* | average hit rate |
| Joint Model | -2325 | 37.8% | -4711 | 13.5% | -9689 | 11.7% |
| Conjoint Model | -2249 | 39.0% | -6333 | 6.8% | -12019 | 7.0% |
| NIC model | -2245 | 39.1% | -5738 | 8.8% | -11100 | 8.6% |
| Aggregate Shares | -2618 | 33.4% | -5830 | 8.4% | -11180 | 8.4% |
| N | 2388 | | 2356 | | 4514 | |

Table 7 summarizes the *lppl* of the conjoint data, the market data used in estimation and the hold-out market data relative to the joint model and our comparison models. Overall, we find that the joint model does a better job than the conjoint model or the NIC model at predicting holdout market choices and thus it is better suited to making predictions about product planning decisions that will play out in the market.

**CONCLUSIONS**

Companies need to accurately predict the effects of product design and marketing mix decisions, and so have traditionally turned to forecasting methods from marketing science. Chief among these are discrete choice models estimated from survey data (e.g., choice-based conjoint). But it is well-known that certain critical quantities can be inaccurately measured by even the most scrupulous conjoint design, for example, reactions to price changes or socially-desirable attributes. Conversely, market data doesn't allow product designers to assess the impact of attributes that are truly new, or do not vary sufficiently among products on the market. The two types of data have complementary strengths, yet prior work attempting to meld them had data requirements so stringent as to render most existing data sources unusable. In this article, we developed a framework for combining survey and individual-level market data originating from separate sources, in a way that adjusts *all* model parameters to be more consistent with purchase behavior observed in the market, without resorting to *post hoc* adjustments that are not part of the statistical estimation process.

Using conjoint and market data for minivans, we found that the model estimated jointly from both conjoint and market data predicts holdout market choices better than one estimated from conjoint data alone, demonstrating the benefits of pooling information from multiple data sources. A particularly useful aspect of the minivan market data was that it included each individual's 'reasons for purchase', which—in contrast to the dominant findings in the empirical modeling literature about demographic variables—turn out to be effective at explaining choice behavior. Using our framework, we were able to incorporate these characteristics, which (along with the conjoint data) help to inform the distribution of heterogeneity for the choice parameters even though we only observe one choice for each individual in the market data. Our

joint model also fits the estimation data better and predicts holdout purchases better than a model which does not include the individual characteristics, demonstrating that including these individual characteristics not only allows for flexibility, but also improves prediction. Such improvements in prediction accuracy are crucial to companies, like GM, that frequently make substantial capital decisions on the basis of such forecasts.

It would be valuable to test this method using other data sets, as the minivan data represented a rather extreme case of missing characteristics. In situations where there is more overlap between the characteristics, we would expect the likelihood-based missing data method to perform even better. The present application also did not allow for a strong test of the ability of the model to accommodate situations where there are observed differences in the distributions of the individual characteristics. Finally, it would be straightforward to apply the method to three or more data sets.

There are a number of extensions to this model that could be considered to accommodate different data. Although we have chosen a logit specification for tractability and to conform to typical conjoint practice, a probit specification could be used if all alternatives were included in each conjoint question. In situations where there is selection bias based on the outcomes, it would also be possible to incorporate a model that accounts for selection on $\beta_n$ (Heckman 1979), such as might happen if the market data contained buyers and the conjoint data sampled both buyers and non-buyers. In durables, it is also common to collect data on the consumer's second choice product and this second-choice could be incorporated in the likelihood. Researchers can readily incorporate such extensions into the Bayesian MCMC sampler.

In the present work, we have also assumed that the parameters that relate an individual's characteristics to his or her attribute preferences are common across choice data sets. However, it is possible that there are systematic differences in choice behavior across different choice contexts. If a large number of choice contexts were observed (say market data across different retailers or conjoint studies fielded at different locations) it would be possible to explicitly model the distribution of $\beta_0$ and $\Delta$ across data sets, adding another level to the hierarchical model. The resulting estimates of $\beta_0$ and $\Delta$ for each choice

context would be optimally shrunk toward the population mean according to the number of individuals observed in each data set. Dominici et al. (1997) applied a similar idea in the context of hierarchical linear models for meta-analysis of regression studies.

We envision uses for the proposed methodology outside product design, and even beyond empirical marketing. Among the most active areas of research in marketing is the analysis of web data. But understanding web behavior is bedeviled by the need to combine data sets across many web sites, product types, and consumer groups. The proposed model provides a framework to handle data spanning these dimensions, so long as a plausible unifying mechanism (relating, say, customer or product characteristics to the coefficient values) could be specified. A more grandiose issue involves the "internal vs. external validity debate" in experimental design. Note that experiments—with their tight controls and potential for orthogonalization—are the gold standard for internal validity, while matching market data is the very definition of external validity. In allowing the analyst to meld these sources, the proposed model offers a platform for capitalizing on both the rigorous checks of internally valid survey data and the empirical fidelity of externally valid market data.

**REFERENCES**

Allenby, Greg M. and James L. Ginter. 1995. Using Extremes to Design Products and Segment Markets. Journal of Marketing Research, 32(4), pp 392-403.

Ben-Akiva, M., M. Bradley, T. Morikawa, J. Benjamin, T. Novak, H. Oppewal, and V. Rao. 1994. Combining Revealed and Stated Preferences Data. Marketing Letters, 5(4), pp 335-350.

Ben-Akiva, Moshe and Takayuki Morikawa. 1990. Estimation of switching models from revealed preferences and stated intentions. Transportation Research A, 24A(6), pp 485-495.

Bhat, Chandra R. and Saul Castelar. 2002. A unified mixed logit framework for modeling revealed and stated preferences: formulation and application to congestion pricing analysis in the San Francisco Bay area. Transportation Research B, 36, pp 593-616.

Blamey, Russell and Jeff Bennett. 2001. Yea-saying and Validation of a Choice Model of Green Product Choice. in Jeff Bennett and Russell Blamey, eds., The Choice Modeling Approach to Environmental Valuation. Edward Elgar, Northampton, MA, pp178-201.

Bradley, Mark and Andrew Daly. 1994. Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. Transportation, (21)2, pp 167-184.

Brooks, Stephen P. and Andrew Gelman. 1998. General Methods for Monitoring Convergence of Iterative Simulations. Journal of Computational and Graphical Statistics, 7(4), pp 434-455.

Brownstone, David, David S. Bunch and Kenneth Train. 2000. Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. Transportation Research B, 34, pp 315-338.

Chib, Siddhartha and Edward Greenberg. 1998. Analysis of Multivariate Probit Models. Biometrika, 85(2), pp 347-361.

De Bruyn, Arnaud, John C. Liechty, Eelko K.R.E. Huizingh and Gary L. Lilien. 2007. Offering Online Recommendations with Minimum Customer Input through Conjoint-Based Decision Aids, Marketing Science, forthcoming.

Dominici, Francesca, Giovanni Parmigiani, Kenneth H. Reckhow and Robert L. Wolpert. 1997. Comgining Information from Related Regressions. Journal of Agricultural, Biological and Environmental Statistics, 2(3), pp 313-332.

Fader, Peter S. and Bruce G.S. Hardie. 1996. Modeling Consumer Choice among SKUs. Journal of Marketing Research**,** 33(4), pp 442-452.

Fennell, Geraldine, Greg M. Allenby, Sha Yang and Yancy Edwards. 2003. The Effectiveness of Demographic and Psychographic Variables for Explaining Brand and Product Category Use. Quantitative Marketing and Economics, 1(2), pp 223-244.

Gilbride, Timothy J., Peter J. Lenk and Jeff D. Brazell. 2006. Market Share Constraints and the Loss Function in Choice Based Conjoint Analysis. Marketing Science, forthcomming.

Green, Paul E., Abba M. Krieger and Yoram (Jerry) Wind. 2001. Thirty Years of Conjoint Analysis: Reflections and Prospects. Interfaces, 31(3), pp S56 – S73.

Green, Paul E. and Vithala R. Rao. 1971. Conjoint Measurement for Quantifying Judgmental Data. Journal of Marketing Research, 8(3), pp 355-363.

Haario, Heikki, Eero Saksman, and Johanna Tamminen. 2001. An adaptive Metropolis algorithm. Bernoulli, 7(2), pp 223-242.

Heckman, James J. 1979. Sample Bias As A Specification Error. Econometrica, 1979, 47(1), pp 153-162.

Ibrahim, Joseph G., Stuart R. Lipsitz and Ming-Hui Chen. 1999. Missing characteristics in generalized linear models when the missing data mechanism is non-ignorable. Journal of the Royal Statistical Society, 61, pp 173-190.

Krishnan, V. and Karl T. Ulrich. 2001. Product Development Decisions: A Review of the Literature. Management Science, 47(1), pp 1-21.

Lenk, Peter J., Wayne S. DeSarbo, Paul E. Green and Martin R. Young. 1996. Hierarchical Bayes Conjoint Analysis: Recovery of Parthworth Heterogeneity from Reduced Experimental Designs. Marketing Science, 15(2), pp 173-191.

Little, Roderick J.A. and Donald B. Rubin. 2002. Statistical Analysis with Missing Data, Second Edition. Wiley Interscience, Hoboken, New Jersey.

Louviere, Jordan J., David A. Hensher and Joffre D. Swait. 2000. Stated Choice Methods: Analysis and Application. Cambridge University Press, Cambridge, UK.

Louviere, Jordan J., Robert J. Meyer, David S. Bunch, Richard Carson, Benedict Dellaert, W. Michael Hanemann, David Hensher and Julie Irwin. 1999. Combining Sources of Preference Data for Modeling Complex Decision Processes. Marketing Letters, 10(3), pp 205-217.

Manski, Charles F., Steven R. Lerman. 1977. The Estimation of Choice Probabilities from Choice Based Samples. Econometrica, (45)8, pp 1977-1988.

McCulloch, Robert and Peter E. Rossi. 1994. An exact likelihood analysis of the multinomial probit model. Journal of Econometrics, 64, pp 207-240.

Newton, Michael A. and Adrian E. Raftery. 1994. Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. Journal of the Royal Statistical Society Series B (Methodological), 56(1), pp 3-48.

Orme, Bryan and Rich Johnson. 2006. External Effects Adjustments in Conjoint Analysis. Sawtooth Software Research Paper Series.

Rossi, Peter E., Greg M. Allenby and Rob McCulloch. 2005. Bayesian Statistics and Marketing. John Wiley and Sons, West Sussex, England.

Sandor, Zsolt and Michel Wedel. 2005. Heterogeneous Conjoint Choice Designs. Journal of Marketing Research, 42(2), pp 210-218.

Swait, Joffre and Rick L. Andrews. 2003. Enriching scanner panel models with choice experiments. Marketing Science, 22(4), pp 442-460.

Swait, Joffre and Jordan Louviere. 1993. The role of the scale parameter in the estimation and comparison of multinomial logit models. Journal of Marketing Research, 30, pp 305-314.

Tanner, Martin A. and Wing Hung Wong. 1987. The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association, 82(398), pp 528-540.

Urban, Glen L., John R. Hauser, John H. Roberts. 1990. Prelaunch Forecasting of New Automobiles. Management Science, 36(4), pp 401-421.

Verlegh, Peeter W. J., Hendrik N.J. Sihfferstein and Dick R. Wittink. 2002. Range and Number-of-Levels Effects in Derived and Stated Measures of Attribute Importance. Marketing Letters, 13(1), pp 41-52.

**E-COMPANION**

This e-companion contains a description of the priors and sampler algorithm used to estimate the joint model (section EC.1), a detailed description of the parameter recovery study (section EC.2) and estimates of the population parameters for the conjoint and No-Individual Characteristics (NIC) models (section EC.3).

## EC.1. Priors & Sampler Algorithm

Throughout section EC.1, we will assume that $z_n$ includes the value 1 as its leading element for all $n$ and that $\Delta$ includes an initial column to multiply these initial ones. Thus the intercept $\beta_0$ is incorporated into $\Delta$.

### Priors

We use proper but diffuse conditionally conjugate priors. Specifically,

$$[\Delta] = \mathrm{MVN}\left(\mathrm{vec}(\Delta) | \mathbf{0}, diag(vec(F))\right)$$

$$[\Sigma_v] = \mathrm{I}W(\Sigma_v \mid K+2, I)$$

$$[\mu_z] = MVN(\mu_z \mid \mathbf{0}, 1000I)$$

$$[\Sigma_z] = IW(\Sigma_z \mid L+2, I)$$

$$[\mu] = \mathrm{Gamma}(\mu \mid 1/1000, 1000)$$

where $I$ is the identity matrix, $K$ is the number of attributes and $L$ is the number of individual characteristics plus one for the intercept. Note that we are not using the more common structured prior for $\Delta$ (c.f. Rossi, Allenby and McCulloch 2005, p. 71). The non-standard form of the prior on $\Delta$ allows us to specify tight priors on particular elements of $\Delta$, which we use to restrict which individual characteristics relate to particular attribute preferences. $F$ is a matrix of the same dimension as $\Delta$ that is used to determine which elements of $\Delta$ have tight priors near zero and which have diffuse priors. For a tight prior, the corresponding element of $F$ had a value of $10^{-12}$. For a diffuse prior, the corresponding element of $F$ had a value of 1000.

**Sampler Algorithm**

**Step 0. Initialize values for $\mu$, $\Delta$, $\Sigma_\upsilon$, $\mu_z$ and $\Sigma_z$, and for $\beta_n$ and $z_n^{mis}$ for all $n$.**

The scale ratio $\mu$ is initialized to the maximum likelihood estimate from a non-homogeneous joint model. The parameters in $\Delta$ are initialized at their maximum likelihood estimates from a homogeneous logit model that includes interactions between elements of $w_n$ (coded -1,1) and $x_{njt}$ and was estimated from the market data. The vector $\mu_z$ is initialized to zero. The matricies $\Sigma_\upsilon$ and $\Sigma_z$ are initialized to identity matrices. Starting values of $z_n$ are drawn based on $\mu_z$, $\Sigma_z$ and any observed $w_n$. Then, we generate starting values for $\beta_n$ and $z_n^{mis}$ according to the model.

**Step 1. For each individual, $n$, draw $\beta_n$.**

$$[\beta_n | \{y_{nt}\}, \{X_{nt}\}, z_n, \mu, \Delta, \Sigma_\nu] \propto \left( \prod_t [y_{nt} | X_{nt}, \beta_n, \mu] \right) [\beta_n | z_n, \Delta, \Sigma_\nu]$$

$$\propto \left( \prod_{t \in \text{data set 1}} \frac{\exp(\beta_n x_{n,y_{nt},t})}{\sum_j \exp(\beta_n x_{njt})} \right) \left( \prod_{t \in \text{data set 2}} \frac{\exp(\mu\beta_n x_{n,y_{nt},t})}{\sum_j \exp(\mu\beta_n x_{njt})} \right) N_K(\beta_n | \Delta z_n, \Sigma_\nu)$$

This distribution is not a standard distribution and we use a Metropolis-Hastings step to complete the draw. The proposal is a multivariate normal random-walk from the most recent draw where the covariance of the random walk for individual $n$ is based on the covariance of all previous draws for individual $n$. Haario, Saksman and Tamminen (2001) show that if all previous draws (not just a window) are used to compute the covariance used in the proposal, the ergodic properties of the chain are preserved. To simplify computation, a recursive formula is used to update the covariance for the proposal with each draw.

**Step 2. Draw $\mu$.**

$$[\mu|y,X,\{\beta_n\}] \propto \left(\prod_{(n,t)\in\text{data set 2}}[y_{nt}\mid X_{nt},\beta_n]\right)[\mu]$$
$$\propto \left(\prod_{t\in\text{data set 2}}\frac{\exp(\mu\beta_n x_{n,y_{nt},t})}{\sum_j \exp(\mu\beta_n x_{njt})}\right)\text{Gamma}(\mu \mid 1/1000, 1000)$$

We make this draw using a Metropolis-Hastings step with a normal random walk proposal. Note that this draw depends only on the choice observations for the conjoint data.

**Step 3. Draw $\Delta$.**

$$[\text{vec}(\Delta)|Z,\{\beta_n\},\Sigma_\nu] \propto \left(\prod_n [\beta_n \mid z_n,\Delta,\Sigma_\nu]\right)[\Delta]$$
$$\propto N_{KL}(\overline{\mu}_\Delta, \overline{\Sigma}_\Delta)$$

where,

$$\overline{\Sigma}_\Delta = \left(Z'Z\otimes\Sigma_\nu^{-1} + \text{diag}(\text{vec}(F))^{-1}\right)^{-1}$$
$$\overline{\mu}_\Delta = \overline{\Sigma}_\Delta\left(\left(Z'\otimes\Sigma_\nu^{-1}\right)\text{vec}(\beta') + \text{diag}(\text{vec}(F))^{-1}(\mathbf{0})'\right)$$

where $Z$ is the matrix obtained by stacking the row vectors $z_n'$ and $\beta$ is the matrix obtained by stacking the row vectors $\beta_n'$.

**Step 4. Draw $\Sigma_\nu$.**

$$[\Sigma_\nu|Z,\{\beta_n\},\Sigma_\nu] \propto \left(\prod_n [\beta_n \mid z_n,\Delta,\Sigma_\nu]\right)[\Sigma_\nu]$$
$$\propto IW(K+2+N,(K+2)I^{-1}+(\beta-Z\Delta')'(\beta-Z\Delta'))$$

Recall that $N$ is the number of individuals in the sample and $K$ is the number of attributes.

**Step 5. For each $n$, draw $z_n$.**

$$[z_n \mid w_n,\mu_z,\Sigma_z] \propto [\beta_n \mid z_n\Delta,\Sigma_\nu][z_n \mid \mu_z,\Sigma_z]$$
$$\propto MVN(\beta_n \mid z_n\Delta,\Sigma_\nu)MVN(z_n \mid \mu_z,\Sigma_z)I(z_n\in B_n)$$
$$= MVN(z_n \mid \tilde{\mu},\tilde{\Sigma})I(z_n\in B_n)$$

$$B_n = B_{n1} \times B_{n2} \times \cdots \times B_{nL} \text{ and } B_{nl} = \begin{cases} (-\infty, 0] & \text{if } w_{nl} = 0 \\ (0, \infty) & \text{if } w_{nl} = 1 \\ (-\infty, \infty) & \text{if } w_{nl} \text{ is missing} \end{cases}$$

$$\widetilde{\mu} = \mu_z + \Sigma_z \widetilde{\Delta}^T \left( \Sigma_\nu + \widetilde{\Delta} \Sigma_z \widetilde{\Delta}^T \right)^{-1} \left( \beta_n - \widetilde{\Delta} \mu_z \right)$$

$$\widetilde{\Sigma} = \Sigma_z - \widetilde{\Delta} \Sigma_z^T \left( \Sigma_\nu + \widetilde{\Delta} \Sigma_z \widetilde{\Delta}^T \right)^{-1} \Sigma_z \widetilde{\Delta}^T$$

where $\widetilde{\Delta}$ is columns 2 through $L$ of $\Delta$. We sample from this truncated normal distribution by sequentially drawing univariate truncated normal Gibbs samples of each element of $z_n$ (McCulloch and Rossi 1994).

**Step 6. Draw $\mu_z$.**

$$\begin{aligned} [\mu_z | Z, \Sigma_z] &\propto \left( \prod_n [z_n \mid \mu_z, \Sigma_z] \right) [\mu_z] \\ &\propto N_L(\mu_z \mid \overline{\mu}_z, \overline{\Sigma}_z) \end{aligned}$$

where,

$$\overline{\Sigma}_z = \left( N \Sigma_z^{-1} + (1000I)^{-1} \right)^{-1}$$

$$\overline{\mu}_z = \overline{\Sigma}_z \left( \Sigma_z^{-1} \sum_n z_n + (1000I)^{-1} \mathbf{0} \right)$$

**Step 7. Draw $\Sigma_z$.**

$$\begin{aligned} [\Sigma_z | Z, \mu_z] &\propto \left( \prod_n [z_n \mid \mu_z, \Sigma_z] \right) [\Sigma_z] \\ &\propto IW(\nu_{\Sigma_z}, S_{\Sigma_z}) \end{aligned}$$

where,

$$\nu_{\Sigma_z} = L + 2 + N$$

$$S_{\Sigma_z} = (L+2)I + \sum_n (z_n - \mu_z)'(z_n - \mu_z)$$

## EC.2 Details of synthetic data generation for parameter recovery study

To understand the parameter recovery properties of the model, we simulated data according to the model. The simulated data set had 1100 individuals: 100 from a hypothetical 'conjoint data' set and 1000

from a hypothetical 'market data' set. We generated a vector of 5 latent continuous characteristics for each individual ($z_n$) according to a multivariate normal distribution with mean $\mu_z = \mathbf{0}$. The variance for each element was 1 and the second and third characteristics had a correlation of 0.4. All other characteristics were independent.

We generated choice parameters ($\beta_n$) for each individual according to the model $\beta_n = \beta_0 + \Delta z_n + \nu_n$, where $z_n$ is the original vector of length 5. The population parameters $\Delta$ and $\Sigma_\nu$ were assumed to be:

$$\beta_0 = \begin{pmatrix} 3 \\ -3 \end{pmatrix}$$

$$\Delta = \begin{pmatrix} -1 & 1 & 1 & 1 & 1 \\ 0.5 & 2 & 0 & 0 & 0 \end{pmatrix}$$

$$\Sigma_\nu = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

For each individual in the conjoint study, we generated 100 choice observations for each individual from choice sets with three alternatives. The product attributes ($x_{njt}$) used in each choice observation were generated from independent standard normals, so each choice observation has a unique set of attribute values. We assumed that the scale ratio between data sets was $\mu = 1.5$ and we generated choice observations for each individual in the market data using the choice parameters $\mu\beta_n$. For the market data, we generated 1 choice observation for each individual from a choice set of size twelve. The product attributes for the market data were also generated from independent standard normals. (Note that this is unlikely to be true in real market data where there are often significant correlations between attributes. In this way, our synthetic market data is more informative than real market data is likely to be.)

To show that we are able to recover the population parameters in circumstances similar to that in the GM minivan data, we ran our estimation algorithm assuming that for the 1000 market respondents, the researcher observed a vector of two binary variables ($w_n$) indicating whether the first two characteristics ($z_{n1}$ and $z_{n2}$) were positive or negative. We assumed the researcher did not observe anything about the other three continuous characteristics. For the conjoint respondents, we assumed that the researcher ob-

served the binary indicator for just the second characteristic. So, for the conjoint respondents, the first binary indicator was 'missing' and is imputed based on the observed choices and the distribution of the covariates in the market data. We used diffuse, but proper priors. Inference was based on 20,000 draws from our MCMC algorithm with a burn-in of 6,000 draws. The draws were thinned to every twentieth draw to reduce data storage. Trace plots indicated that the chain had clearly converged after 6,000 draws. Table EC.1 shows that the recovery of the population level parameters is quite good for this base case.

**Table EC.1. Recovery of population-level parameters ($\Delta$, $\Sigma_v$, $\mu_z$ and $\Sigma_z$)**

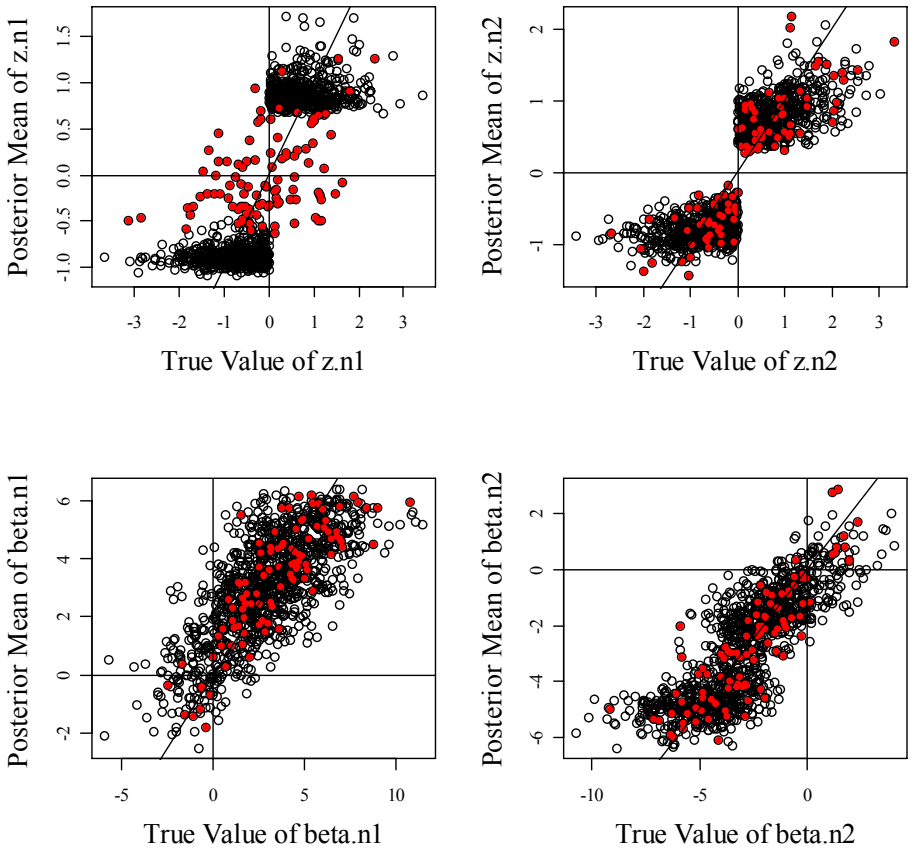
| Parameter | True Value | Posterior Mean | Posterior SD |
|---|---|---|---|
| Delta.11 | 3.00 | 3.07 | 0.18 |
| Delta.21 | -3.00 | -3.02 | 0.17 |
| Delta.12 | -1.00 | -1.08 | 0.16 |
| Delta.22 | 0.50 | 0.54 | 0.15 |
| Delta.13 | 1.00 | 1.34 | 0.16 |
| Delta.23 | 2.00 | 2.23 | 0.16 |
| mu | 0.75 | 0.74 | 0.06 |
| Sigma.11 | 4.00 | 3.60 | 0.62 |
| Sigma.12 | 0.50 | 0.44 | 0.35 |
| Sigma.22 | 1.00 | 0.83 | 0.27 |
| mu.w.1 | 0.00 | 0.01 | 0.04 |
| mu.w.2 | 0.00 | 0.03 | 0.04 |

To understand recovery of the individual-level parameters, $z_n$ and $\beta_n$, we computed the posterior means of these parameters for each individual. Figure EC.1 shows a plot of these posterior means against the true values that were used to generate the data. The closed red circles represent conjoint individual and the open circles represent market individuals. The top panels in Figure EC.1 show recovery of $z_n$. The binary indicator for the second characteristic is observed for all individuals, so the model always predicts the correct sign for $z_{n1}$. For the first characteristic, $z_{n2}$, which is not observed for conjoint individuals, there are a number of conjoint individuals for which the mean of the posterior distribution is not the same sign as the true value.

The bottom panels in Figure EC.1 shows the recovery of $\beta_n$. For the conjoint individuals (indicated with red circles) the posterior mean is a very good estimate of the true choice parameters. However,

we are also able to get reasonably good recovery of the individual-level parameters for the market respondents (open circles), even with just one choice observation per respondent.

**Figure EC.1. Recovery of individual-level characteristics ($z_n$) and choice parameters ($\beta_n$) across all respondents.**



## EC.3 Parameter Estimates for Alternative Model Specifications

Below we provide the posterior means of the population parameters for the conjoint and the No-Individual-Characteristics models which are presented as alternatives to the joint model in Table 7.

### Table EC.2. Estimated parameters for conjoint model.

| | | ***Delta*** | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Other Attributes (X) | | | | Brands (X) | | | | | | | | | | | | | |
| | | Styling Appeal | Price (linear) | Price (squared) | Design Age | A | B | C | D | E | F | G | H | I | J | K | L | M | ***mu.z*** |
| ***Cov*** | Intercept | **1.38** | **-1.06** | **-0.31** | | **0.76** | **0.65** | 0.19 | -0.09 | **1.26** | **-1.61** | -0.26 | -0.15 | -0.24 | 0.23 | **-0.98** | **1.11** | **-1.04** | NA |
| | Household.with.Children | | -0.05 | -0.01 | | | | | | | | | | | | | | | -0.06 |
| ***Sigma.nu*** | | **1.14** | **2.26** | **0.45** | | **2.45** | **3.43** | **2.76** | **2.03** | **3.93** | **2.76** | **1.89** | **2.28** | **3.34** | **2.31** | **3.06** | **6.21** | **6.31** | |

* Values in boldface have a posterior mean more than two posterior standard errors different than zero.

### Table EC.3 Estimated parameters for No Individual Characteristics formulation.

| ***mu*** | *Delta* | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.76 | Other Attributes (X) | | | | Brands (X) | | | | | | | | | | | | | |
| | Styling Appeal | Price (linear) | Price (squared) | Design Age | A | B | C | D | E | F | G | H | I | J | K | L | M | |
| *Intercept* | **1.86** | **-1.18** | **-0.23** | **1.07** | **1.27** | **1.18** | 0.23 | -0.10 | **1.55** | **-1.94** | **-1.01** | -0.22 | -0.21 | 0.15 | -0.51 | **0.94** | **-1.43** | |
| *Sigma.nu* | **2.03** | **3.88** | **0.71** | **4.54** | **6.12** | **6.72** | **4.26** | **2.51** | **8.08** | **6.59** | **4.06** | **2.51** | **5.06** | **2.44** | **5.24** | **7.38** | **6.89** | |

* Values in boldface have a posterior mean more than two posterior standard errors different than zero.