Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Modeling Multimodal Continuous Heterogeneity in Conjoint Analysis—A Sparse Learning Approach

Yupeng Chen, Raghuram Iyengar, Garud Iyengar

To cite this article:
Yupeng Chen, Raghuram Iyengar, Garud Iyengar (2017) Modeling Multimodal Continuous Heterogeneity in Conjoint Analysis—A Sparse Learning Approach. Marketing Science 36(1):140-156. http://dx.doi.org/10.1287/mksc.2016.0992

http://pubsonline.informs.org/journal/mksc/

MARKETING SCIENCE
Vol. 36, No. 1, January–February 2017, pp. 140–156
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Modeling Multimodal Continuous Heterogeneity in Conjoint Analysis—A Sparse Learning Approach

Yupeng Chen,[a] Raghuram Iyengar,[a] Garud Iyengar [b]

[a] Marketing Department, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [b] Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027

Contact: yupengc@wharton.upenn.edu (YC); riyengar@wharton.upenn.edu (RI); garud@ieor.columbia.edu (GI)

Received: July 12, 2013
Accepted: December 22, 2015
Published Online in Articles in Advance: July 19, 2016

https://doi.org/10.1287/mksc.2016.0992

Abstract. Consumers' preferences can often be represented using a multimodal continuous heterogeneity distribution. One explanation for such a preference distribution is that consumers belong to a few distinct segments, with preferences of consumers in each segment being heterogeneous and unimodal. We propose an innovative approach for modeling such multimodal distributions that builds on recent advances in sparse learning and optimization. We apply the model to conjoint analysis where consumer heterogeneity plays a critical role in determining optimal marketing decisions. Our approach uses a two-stage divide-and-conquer framework, where we first divide the consumer population into segments by recovering a set of candidate segmentations using sparsity modeling, and then use each candidate segmentation to develop a set of individual-level heterogeneity representations. We select the optimal individual-level heterogeneity representation using cross-validation. Using extensive simulation experiments and three field data sets, we show the superior performance of our sparse learning model compared to benchmark models including the finite mixture model and the Bayesian normal component mixture model.

History: Preyas Desai served as the editor-in-chief and Joel Huber served as associate editor for this article.
Supplemental Material: Data and the online appendix are available at https://doi.org/10.1287/mksc.2016.0992.

Keywords: sparse machine learning • multimodal continuous heterogeneity • conjoint analysis

1. Introduction

Marketing researchers and practitioners frequently use conjoint analysis to recover consumers' heterogeneous preferences (Green and Srinivasan 1990, Wittink and Cattin 1989), which serve as a critical input for many important marketing decisions, such as market segmentation (Vriens et al. 1996) and differentiated product offerings and pricing (Allenby and Rossi 1998). In practice, consumer preferences can often be modeled using a multimodal continuous heterogeneity (MCH) distribution, where the consumer population is interpreted as consisting of a few distinct segments, each of which contains a heterogeneous subpopulation. Since in most conjoint applications researchers use short questionnaires because of concerns over response rates and response quality, the amount of information elicited from each respondent is limited; therefore, adequate modeling of MCH becomes critical.

Modeling MCH raises two major challenges. First, both across-segment and within-segment heterogeneity must be accommodated to fully capture preference variations among consumers. Second, when pooling data across respondents it is important to impose an adequate amount of shrinkage to recover the individual-level partworths. The widely used finite mixture (FM) model approximates MCH using discrete mass points, each representing a segment of homogeneous consumers (Kamakura and Russell 1989, Chintagunta et al. 1991). While such a discrete representation of the heterogeneity distribution accommodates across-segment heterogeneity, it does not allow for within-segment heterogeneity. Hierarchical Bayes (HB) models with flexible parametric specifications for the heterogeneity distribution have also been proposed to model MCH. For instance, Allenby et al. (1998) developed a Bayesian normal component mixture (NCM) model in which a mixture of multivariate normal distributions is utilized to represent consumers' heterogeneous preferences. While the NCM model is capable of modeling a variety of heterogeneity distributions, it may not be able to impose an adequate amount of shrinkage to accurately recover the individual-level partworths (Evgeniou et al. 2007). Additionally, it faces inferential challenges when conducting a segment-level analysis (Rossi et al. 2005), including the label switching problem (Celeux et al. 2000, Stephens 2000) and the overlapping mixtures problem (Kim et al. 2004).[1]

In this paper, we propose an innovative sparse learning (SL) approach to address both challenges in modeling MCH and apply it in the context of metric and choice-based conjoint (CBC) analysis. Our SL approach models MCH using a two-stage divide-and-conquer framework. In the first stage, we build on recent advances in sparse learning (Tibshirani 1996, Yuan and Lin 2005, Argyriou et al. 2008) to "divide" the MCH distribution and recover a set of candidate segmentations of the consumer population. We make a simple observation that any two respondents from the same segment have identical segment-level partworths. Suppose the population is comprised of a few distinct segments. Then, a substantial proportion of pairwise differences of respondents' segment-level partworths will be zero vectors; in other words, the pairwise differences of respondents' segment-level partworths will be sparse. Our model leverages this observation and learns the sparsity pattern from the conjoint data to recover informative segmentations of the consumer population. In the second stage, we use each candidate segmentation to develop a set of individual-level representations of MCH by separately "conquering" the within-segment heterogeneity distribution of each segment. In particular, for each segment, we model its within-segment heterogeneity assuming a unimodal continuous heterogeneity (UCH) distribution, which is considerably easier to model compared to MCH. We select the optimal individual-level representation of MCH using cross-validation (Wahba 1990, Shao 1993, Vapnik 1998, Hastie et al. 2001). Using the two-stage framework, our SL model accounts for both across-segment and within-segment heterogeneity, and is able to endogenously select an adequate amount of shrinkage for recovering the individual-level partworths. Moreover, since our SL model automatically generates a segmentation of the consumer population, a segment-level analysis can be readily conducted.

We add to the growing literature of machine learning-based methods for conjoint estimation (Toubia et al. 2003, 2004; Evgeniou et al. 2005; Cui and Curry 2005; Evgeniou et al. 2007). This stream of research has largely ignored consumer heterogeneity, with the exception of Evgeniou et al. (2007), who proposed a convex optimization (CO) model for capturing UCH. Our work contributes by developing the first machine learning-based approach to modeling the more general MCH.

We compare our SL model to the FM model, the NCM model, and the CO model using extensive simulation experiments and three field data sets. In simulations, the SL model shows a consistently strong performance in terms of both parameter recovery and predictive accuracy across a wide range of experimental conditions. The results from the simulations shed light on when and why the SL model outperforms other benchmarks. For instance, the performance of the NCM model relative to the SL model is weak when the within-segment variance is small or when the amount of respondent-level data is limited. The latter highlights the usefulness of our approach in contexts where researchers prefer to elicit consumer preferences using short conjoint questionnaires due to concerns over response rates and response quality (Lenk et al. 1996). This pattern of results happens largely because the amount of shrinkage imposed by the NCM model is influenced by exogenously chosen parameters for the second-stage priors and can be inadequate depending on the characteristics of a conjoint data set. In field data, the SL model also shows strong performance in terms of predictive accuracy, and its estimates of individual-level partworths display shapes consistent with MCH. Moreover, in an optimal pricing exercise, the SL model generates a more plausible revenue-maximizing price compared to that from other benchmarks, showing the managerial relevance of using our approach to model MCH in conjoint analysis.

The remainder of this paper is organized as follows. In Section 2 we present our SL model for modeling MCH in conjoint analysis. We compare the SL model and the benchmark methods using simulation experiments in Section 3 and three field conjoint data sets in Section 4. We conclude in Section 5.

2. Model

In this section, we present our SL approach to model MCH in conjoint analysis. Specifically, we give a detailed description of our approach in the context of metric conjoint analysis. We discuss the modifications needed for choice-based conjoint analysis in the Web appendix.

2.1. Metric Conjoint Setup

We assume a total of I consumers (or respondents), each rating J profiles with p attributes. Let the $1 \times p$ row vector x_{ij} represent the jth profile rated by the ith respondent, for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$, and denote by $X_i \triangleq [\, x_{i1}^\top, x_{i2}^\top, \ldots, x_{iJ}^\top \,]^\top$ the $J \times p$ design matrix for the ith respondent. For respondent i, the $p \times 1$ column vector β_i is used to denote her partworths, and her ratings are contained in the $J \times 1$ column vector $Y_i \triangleq (y_{i1}, y_{i2}, \ldots, y_{iJ})^\top$. We assume additive utility functions, i.e., $Y_i = X_i\beta_i + \epsilon_i$, for $i = 1, 2, \ldots, I$, where ϵ_i denotes the random error. The additive specification of the utility functions is a standard assumption in the conjoint analysis literature (Green and Srinivasan 1990).

2.2. Model Overview

Under a MCH distribution, the consumer population is interpreted as consisting of a few distinct

segments of heterogeneous consumers. To fully capture such a heterogeneity structure, a model needs to be sufficiently flexible to accommodate both across-segment and within-segment heterogeneity. It is also critical that the model has the capacity to impose an adequate amount of shrinkage when recovering the individual-level partworths. These considerations motivate a divide-and-conquer strategy for modeling MCH, where the MCH distribution is "divided" into a collection of within-segment UCH distributions, and each UCH distribution is separately "conquered" using established estimation methodologies. We implement this modeling strategy using the following two-stage framework.

In the first stage, we develop a novel sparse learning model to divide the MCH distribution and recover a set of candidate segmentations of the consumer population. Our model is built on the simple observation that any two respondents from the same segment may have different individual-level partworths but must share identical segment-level partworths; i.e., the difference between their respective segment-level partworths is the zero vector. Since the consumer population consists of a few distinct segments, a substantial proportion of pairwise differences of respondents' segment-level partworths are zero vectors; in other words, the pairwise differences of respondents' segment-level partworths are sparse. Leveraging this observation, we use the sparse learning model to learn such sparsity patterns from conjoint data and recover informative candidate segmentations of the consumer population. Each candidate segmentation provides a decomposition of the MCH distribution into a collection of within-segment heterogeneity distributions, which we utilize in the second stage.

In the second stage, we use each candidate segmentation to develop a set of individual-level representations of MCH. Given a candidate segmentation, we separately model the within-segment heterogeneity distribution of each segment assuming a UCH distribution. UCH provides a reasonable characterization of the within-segment heterogeneity distributions and is considerably easier to model than MCH. We choose the CO model of Evgeniou et al. (2007) to model the within-segment distributions, which allows for an effective approach to control the amount of shrinkage imposed when modeling UCH. We select the optimal individual-level representation of MCH using cross-validation (Wahba 1990, Shao 1993, Vapnik 1998, Hastie et al. 2001). The cross-validation procedure provides a fully data-driven approach to endogenously select an adequate candidate segmentation and an adequate amount of shrinkage to recover the individual-level partworths.

2.3. First Stage: Recovering Candidate Segmentations

The first stage of our SL model aims at learning a set of candidate segmentations of the MCH distribution. To motivate, we consider a standard characterization of the data-generating process of MCH (Andrews et al. 2002a, b). The data-generating process selects the number of segments L, the segment-level partworths $\{\hat{\beta}_l^S\}_{l=1}^L$, and the segment-membership matrix $Q \in \mathbb{R}^{I\times L}$, where $Q_{il} = 1$ if respondent i is assigned to segment l, and $Q_{il} = 0$ otherwise. If respondent i belongs to segment l, she receives a copy of segment-level partworths $\beta_i^S = \hat{\beta}_l^S$, and her individual-level partworths are determined by $\beta_i = \beta_i^S + \xi_i$, where ξ_i denotes the difference between respondent i's segment-level and individual-level partworths, i.e., the within-segment heterogeneity. Let $\hat{B}^S \triangleq \{\hat{\beta}_l^S\}_{l=1}^L$, $B^S \triangleq \{\beta_i^S\}_{i=1}^I$, and $B \triangleq \{\beta_i\}_{i=1}^I$.

Assuming the above data-generating process, recovering candidate segmentations can be achieved by learning the set of model parameters $\{L, \hat{B}^S, Q, B^S, B\}$ from the conjoint data. A closer examination reveals that learning $\{B^S, B\}$ is sufficient, as other model parameters $\{L, \hat{B}^S, Q\}$ can be uniquely determined from $\{B^S, B\}$. We highlight the following three assumptions about the data-generating process that are relevant to learning $\{B^S, B\}$:

Assumption 1 (A1). *The ratings vector Y_i is generated based on β_i, i.e., $Y_i = X_i\beta_i + \epsilon_i$.*

Assumption 2 (A2). *The individual-level partworths β_i is generated based on the segment-level partworths β_i^S, i.e., $\beta_i = \beta_i^S + \xi_i$.*

Assumption 3 (A3). *Respondents i and k belong to the same segment if and only if $\beta_i^S - \beta_k^S = 0$.*

Within an optimization framework with $\{B^S, B\}$ as decision variables, A1 (respectively, A2) suggests to penalize the discrepancy between Y_i and $X_i\beta_i$ (respectively, the discrepancy between β_i and β_i^S). A3, together with the observation that for a substantial proportion of $i-k$ pairs, respondents i and k belong to the same segment, implies that the pairwise discrepancies of the true B^S are sparse. It thus suggests that we can impose a sparse structure on the pairwise discrepancies of B^S when learning $\{B^S, B\}$ and use the sparsity pattern to learn the underlying segmentation.

Motivated by these considerations, we propose the following sparse learning problem to recover candidate segmentations.

(Metric-SEG)

$$\min \left\{ \sum_{i=1}^{I} \|Y_i - X_i\beta_i\|_2^2 + \gamma \sum_{i=1}^{I} (\beta_i - \beta_i^S)^\top D^{-1} (\beta_i - \beta_i^S) + \lambda \sum_{1\le i<k\le I} \theta_{ik} \|\beta_i^S - \beta_k^S\|_2 \right\} \quad (1)$$

s.t. D is a positive semidefinite matrix scaled to have trace 1,
$\beta_i, \beta_i^S \in \mathbb{R}^p, \quad \text{for } i = 1, 2, \ldots, I,$

where γ, λ, and $\{\theta_{ik}\}$ are the regularization parameters that control the relative strength of each penalty term in (Metric-SEG). We will discuss the specification of the regularization parameters in a few paragraphs.

In (Metric-SEG), the first two penalty terms are standard quadratic functions measuring the discrepancy between Y_i and $X_i\beta_i$ and that between β_i and β_i^S, respectively. We note that the matrix D is a decision variable and is related to the covariance matrix of the partworths within each segment (Evgeniou et al. 2007). The third penalty term aims to impose the sparse structure suggested by A3 and is the key to the formulation of (Metric-SEG). In particular, it aims to learn whether respondents i and k belong to the same segment by penalizing the ℓ_2-norm of $\beta_i^S - \beta_k^S$, i.e., $\|\beta_i^S - \beta_k^S\|_2$, for all i–k pairs. We choose the ℓ_2-norm to measure the discrepancy between β_i^S and β_k^S since, unlike most standard measures of magnitude of vectors, e.g., the sum-of-squares measure, the ℓ_2-norm is a *sparsity-inducing penalty function* in that it is capable of enforcing *exact* zero value in optimal solutions under a suitable level of penalty.[2] Sparsity-inducing penalty functions play a fundamental role in sparse learning (Tibshirani 1996, Yuan and Lin 2005, Bach et al. 2011). Our use of the ℓ_2-norm to penalize the pairwise differences of B^S can be viewed as a generalization of the overlapping ℓ_1/ℓ_2-norm (Jenatton et al. 2012, Kim and Xing 2012) and the fused lasso penalty (Tibshirani et al. 2004), and was recently introduced in the context of unsupervised learning (Hocking et al. 2011).

The rationale for assessing whether respondents i and k belong to the same segment by penalizing the ℓ_2-norm of $\beta_i^S - \beta_k^S$ is as follows. For the purpose of illustration, suppose we set $\theta_{ik} = 1$ for all i–k pairs in (Metric-SEG), and thus homogenize the penalty imposed on the ℓ_2-norm of $\beta_i^S - \beta_k^S$. For any two respondents i and k, we consider the following components of the objective function of (Metric-SEG):

$$G_{i,k} \triangleq \sum_{r=i,k} \|Y_r - X_r\beta_r\|_2^2 + \gamma \sum_{r=i,k} (\beta_r - \beta_r^S)^\top D^{-1} (\beta_r - \beta_r^S) + \lambda \|\beta_i^S - \beta_k^S\|_2.$$

Within an optimization framework, the three penalty terms in $G_{i,k}$ induce competing shrinkage over the decision variables $\{\beta_r, \beta_r^S\}_{r=i,k}$: the first term shrinks β_r toward the true individual-level partworths $\beta_r(T)$, and the second term shrinks β_r and β_r^S toward each other, for $r = i, k$, whereas the third term shrinks β_i^S and β_k^S toward each other. Whether $\beta_i^S - \beta_k^S = 0$ holds in the optimal solution is largely determined by the trade-off among the three competing shrinkages, which is, in turn, determined by the distance between $\beta_i(T)$ and $\beta_k(T)$ as well as the regularization parameters γ and λ. If respondents i and k are from the same segment, the distance between $\beta_i(T)$ and $\beta_k(T)$ is likely to be small, and a moderate penalty imposed on $\|\beta_i^S - \beta_k^S\|_2$, i.e., a small λ, should be sufficient to enforce $\beta_i^S - \beta_k^S = 0$ due to the sparsity-inducing property of the ℓ_2-norm. If respondents i and k are from distinct segments, the distance between $\beta_i(T)$ and $\beta_k(T)$ is likely to be large, and enforcing $\beta_i^S - \beta_k^S = 0$ can only be achieved when a strong penalty is imposed on $\|\beta_i^S - \beta_k^S\|_2$, i.e., a large λ is specified. This suggests that if γ and particularly λ are appropriately specified, it is possible to recover the underlying segmentation of the consumer population by solving (Metric-SEG) and identifying i–k pairs with $\beta_i^S - \beta_k^S = 0$ in the optimal solution.

Regularization Parameters. We first discuss the specification for the regularization parameters $\{\theta_{ik}\}$. A heterogeneous specification for $\{\theta_{ik}\}$ is useful for (Metric-SEG) because it allows us to incorporate information that could potentially facilitate the recovery of the underlying segmentation. For example, suppose there is information suggesting that the pair of respondents i and k are more likely to be drawn from the same segment compared to the pair of respondents i' and k'. This information can be accommodated in (Metric-SEG) by setting $\theta_{ik} > \theta_{i'k'}$ such that a stronger sparsity-inducing penalty is imposed to enforce $\beta_i^S - \beta_k^S = 0$.

In this paper, we specify $\{\theta_{ik}\}$ as follows:

$$\theta_{ik} = R(W(\bar{\beta}_i, \bar{\beta}_k)), \tag{2}$$

where $\{\bar{\beta}_i\}_{i=1}^I$ are some initial estimates of the individual-level partworths, $W(\cdot, \cdot)$ is a distance measure of two vectors, and $R(\cdot)$ is a positive, nonincreasing function. The rationale for this specification is that when the distance between the initial individual-level partworth estimates $\bar{\beta}_i$ and $\bar{\beta}_k$ is small, it is likely that respondents i and k belong to the same segment, and therefore, θ_{ik} is set to a large value to induce $\beta_i^S - \beta_k^S = 0$. The admissible choices for $\{\bar{\beta}_i\}_{i=1}^I$, $W(\cdot, \cdot)$, and $R(\cdot)$ are quite flexible. In the empirical implementation of our SL model, we choose to estimate $\{\bar{\beta}_i\}_{i=1}^I$ using the CO model of Evgeniou et al. (2007). We set $W(x, y) = ((x - y)^\top \bar{D}^{-1}(x - y))^{1/2}$, where $\bar{D}$ is the scaled covariance matrix of the partworths generated by the CO model along with $\{\bar{\beta}_i\}_{i=1}^I$ (Evgeniou et al. 2007); such a specification gives more weight to difference between two initial individual-level partworth estimates along directions in which there is less variation across respondents. We set $R(x) = e^{-\omega x}$, a positive, nonincreasing function parameterized by a regularization parameter $\omega \geq 0$. Consequently, we adopt the following specification for $\{\theta_{ik}\}$:[3,4]

$$\theta_{ik} = e^{-\omega((\bar{\beta}_i - \bar{\beta}_k)^\top \bar{D}^{-1}(\bar{\beta}_i - \bar{\beta}_k))^{1/2}}. \tag{3}$$

In this specification, the regularization parameter ω controls the extent to which $\{\bar{\beta}_i\}_{i=1}^I$ are used to facilitate recovering candidate segmentations. When $\omega = 0$, $\{\bar{\beta}_i\}_{i=1}^I$ do not enter the specification of $\{\theta_{ik}\}$, and a homogeneous penalty is imposed on the pairwise discrepancies of B^S; as ω increases, $\{\theta_{ik}\}$ become more heterogeneous, and pairs of respondents with closer initial estimates, i.e., those deemed as more likely to be drawn from the same segment, are penalized more heavily than those with farther initial estimates.

Given the specification of $\{\theta_{ik}\}$ in (3), the regularization parameters for (Metric-SEG) are now given by the vector $\Gamma \triangleq (\gamma, \lambda, \omega)$. Since an appropriate value for Γ is not known a priori, we specify a finite grid $\Theta \subset \mathbb{R}^3$ and solve (Metric-SEG) for each $\Gamma \in \Theta$.[5] We denote $(B(\Gamma), B^S(\Gamma), D(\Gamma))$ as the optimal solution of (Metric-SEG) given Γ. For each Γ, we use $B^S(\Gamma)$ to recover a candidate segmentation $Q(\Gamma)$.

Solution Algorithm. (Metric-SEG) is a convex optimization problem for all regularization parameters $\Gamma \in \Theta$, which implies that it is efficiently solvable to global optimum in theory (Boyd and Vandenberghe 2004). However, solving (Metric-SEG) poses an algorithmic challenge since the third penalty term, $\lambda \sum_{1 \le i < k \le I} \theta_{ik} \|\beta_i^S - \beta_k^S\|_2$, is a nondifferentiable and nonseparable function. Nondifferentiability implies that standard convex optimization methods requiring a differentiable objective function, e.g., Newton's method, cannot be applied to solve (Metric-SEG); nonseparability also adds to the complexity (Chen et al. 2012). We solve (Metric-SEG) using a special purpose algorithm based on variable splitting and the alternating direction augmented Lagrangian method that was proposed in Qin and Goldfarb (2012). This algorithm is specifically designed for handling complex sparsity-inducing penalty functions and is capable of solving for the global optimum of (Metric-SEG). We provide a detailed description of the algorithm in the Web appendix.

Dealing with Small Segments. In many instances of (Metric-SEG) encountered in our simulation experiments and field applications, we observed that the candidate segmentation Q contains a small number of substantive segments that comprise the majority of the consumer population, as well as a few segments each consisting of very few respondents, often one or two. Since these small segments bear little practical interpretation, we employ a simple procedure to combine each of the small segments with its closest substantive segment. Formally, we define a segment in Q as a valid segment if it contains at least M respondents, where M is a prespecified threshold, and as an invalid segment otherwise. Without loss of generality, we assume that the first $\bar{L}$ segments of Q are valid. We retain all valid segments, and for each invalid segment, i.e., the lth segment with $l > \bar{L}$, we determine its closest valid segment by computing $c(l) \triangleq \{v \in \{1, 2, \ldots, \bar{L}\} \mid \|\hat{\beta}_v^S - \hat{\beta}_l^S\|_2 < \|\hat{\beta}_{v'}^S - \hat{\beta}_l^S\|_2, \text{ for } v' \in \{1, 2, \ldots, \bar{L}\}, v' \neq v\}$, and combine the lth segment (an invalid segment) and the $c(l)$th segment (a valid segment). We define $\bar{Q}$, the segmentation obtained after this processing, as the candidate segmentation, but still refer to it using Q for simplicity hereafter.[6] We note that it is possible that no valid segment exists in a segmentation, i.e., $\bar{L} = 0$. In such a case, we simply claim that no candidate segmentation is identified for this instance of (Metric-SEG).

Summary. The first stage of our SL model recovers a set of candidate segmentations in the following manner. We specify a finite grid $\Theta \subset \mathbb{R}^3$ from which the regularization parameters $\Gamma = (\gamma, \lambda, \omega)$ are chosen. For each $\Gamma \in \Theta$, we solve (Metric-SEG) and obtain the candidate segmentation $Q(\Gamma)$; $Q(\Gamma)$ could be an empty matrix in cases where no candidate segmentation is identified. We also include the trivial segmentation where all respondents are in one segment as a candidate segmentation, i.e., $Q(\text{Trivial}) \triangleq 1_{I \times 1}$. We denote the set of candidate segmentations as Φ, i.e., $\Phi \triangleq \{Q(\Gamma)\}_{\Gamma \in \Theta: Q(\Gamma) \neq \varnothing} \cup \{Q(\text{Trivial})\}$; Φ is the output of the first stage of the SL model.[7]

2.4. Second Stage: Recovering Individual-Level Partworths

The second stage of our SL model aims at leveraging the set of candidate segmentations Φ to accurately recover the individual-level partworths. To this end, we develop a set of individual-level representations of MCH based on each candidate segmentation and select the optimal individual-level representation of MCH using cross-validation.

Given $Q \in \Phi$, we propose to model MCH by separately modeling the within-segment heterogeneity distribution for each segment assuming a UCH distribution; that is, Q is interpreted as a decomposition of the MCH distribution into a collection of UCH distributions that are considerably easier to model. There are many effective approaches for modeling UCH in the marketing literature, including the unimodal HB models (Lenk et al. 1996, Rossi et al. 1996) and RR-Het, the metric version of the CO model of Evgeniou et al. (2007). We choose RR-Het to model within-segment UCH distributions because it outperforms standard unimodal HB models (Evgeniou et al. 2007) and allows for a direct and parsimonious way for controlling the amount of shrinkage imposed on the individual-level partworth estimates that can be readily incorporated in a cross-validation framework for endogenously selecting an adequate amount of shrinkage.

Formally, for a candidate segmentation Q with L segments, we define a set of *modeling strategies* $\{S \mid S \triangleq (Q, \psi, COV)\}$, parameterized by $\psi = (\psi^1, \psi^2, \ldots, \psi^L)$ and $COV = (COV^1, COV^2, \ldots, COV^L)$, where $\psi^l > 0$ and

$COV^l \in \{\text{General}(G), \text{Restrictive}(R)\}$ for $l = 1, 2, \ldots, L$. The modeling strategy S models MCH and obtains the individual-level partworth estimates $\{\tilde{\beta}_i\}_{i=1}^I$ by solving a convex optimization problem Metric-HET$(Q; l; \psi^l; COV^l)$ for the lth segment of Q, denoted as $\Upsilon(Q; l)$, for $l = 1, 2, \ldots, L$. When $COV^l = G$, the optimization problem (Metric-HET$(Q; l; \psi^l; G)$) is defined as follows:

$$\begin{aligned}&(\text{Metric-HET}(Q; l; \psi^l; G))\\ &\min \left\{ \sum_{i \in \Upsilon(Q;l)} \|Y_i - X_i \tilde{\beta}_i\|_2^2 \right.\\ &\qquad \left. + \psi^l \sum_{i \in \Upsilon(Q;l)} (\tilde{\beta}_i - \tilde{\beta}_0^l)^\top (D^l)^{-1} (\tilde{\beta}_i - \tilde{\beta}_0^l) \right\} \qquad (4)\\ &\text{s.t. } D^l \text{ is a positive semidefinite matrix}\\ &\qquad \text{scaled to have trace 1,}\\ &\qquad \tilde{\beta}_i \in \mathbb{R}^p, \text{ for } i \in \Upsilon(Q; l); \quad \tilde{\beta}_0^l \in \mathbb{R}^p.\end{aligned}$$

When $COV^l = R$, the optimization problem Metric-HET$(Q; l; \psi^l; R)$ is defined as follows:

$$\begin{aligned}&(\text{Metric-HET}(Q; l; \psi^l; R))\\ &\min \left\{ \sum_{i \in \Upsilon(Q;l)} \|Y_i - X_i \tilde{\beta}_i\|_2^2 \right.\\ &\qquad \left. + \psi^l \sum_{i \in \Upsilon(Q;l)} (\tilde{\beta}_i - \tilde{\beta}_0^l)^\top (I/p)^{-1} (\tilde{\beta}_i - \tilde{\beta}_0^l) \right\} \qquad (5)\\ &\text{s.t. } \tilde{\beta}_i \in \mathbb{R}^p, \text{ for } i \in \Upsilon(Q; l); \quad \tilde{\beta}_0^l \in \mathbb{R}^p.\end{aligned}$$

We note that (5) is obtained from (4) by restricting the decision variable $D^l = I/p$. In both optimization problems, the regularization parameter ψ^l provides a direct and parsimonious way to control the trade-off between fit and shrinkage. In particular, a larger ψ^l imposes more shrinkage on the individual-level partworth estimates in the lth segment toward $\tilde{\beta}_0^l$, which can be shown to be the segment mean (Evgeniou et al. 2007), and hence results in more homogenous estimates. The matrix D^l in (4) is related to the covariance matrix of the partworths within the lth segment (Evgeniou et al. 2007). Explicitly modeling D^l allows for a general covariance structure and gives rise to much flexibility in modeling within-segment heterogeneity. On the other hand, restricting $D^l = I/p$ in (5) imposes a restrictive covariance structure that is less flexible but is also more parsimonious and robust with respect to overfitting. We assess the relative strength of the two optimization problems with different covariance structures using cross-validation.

We note that each modeling strategy $S = (Q, \psi, COV)$ gives rise to a distinct individual-level representation of MCH. In particular, the segmentation Q determines the way in which MCH is decomposed into a collection of UCHs, and ψ and COV control the amount of shrinkage imposed and the covariance structure assumed when modeling UCH for each segment of Q, respectively.

Cross-validation. To endogenously select the optimal modeling strategy (and hence the optimal individual-level representation of MCH it implies), we evaluate the *cross-validation error* of each modeling strategy S. Cross-validation is a standard technique used in the statistics and machine learning literature for model selection (Wahba 1990, Shao 1993, Vapnik 1998, Hastie et al. 2001) and has been adopted in the recent literature of machine learning and optimization-based methods for conjoint estimation (Evgeniou et al. 2005, 2007). We measure the cross-validation error of a modeling strategy S, $CVE(S)$, identically as in Evgeniou et al. (2005, 2007). The cross-validation error $CVE(S)$ provides an effective estimate of the predictive accuracy of the modeling strategy S on *out-of-sample* data using only *in-sample* data, i.e., the data available to the researcher for model calibration. To implement cross-validation, we prespecify a finite grid $\Xi \subset \mathbb{R}$, and for each Q we consider modeling strategies $S = (Q, \psi, COV)$ such that $\psi^l \in \Xi$ and $COV^l \in \{G, R\}$, for $l = 1, 2, \ldots, L$.[8] We select S that minimizes $CVE(S)$ as the optimal modeling strategy and its corresponding Q as the optimal candidate segmentation, which we denote by S^* and Q^*, respectively. Consequently, the cross-validation procedure allows us to endogenously select the modeling strategy S^* that is expected to have the optimal predictive accuracy on out-of-sample data. We recover the optimal individual-level partworth estimates $\{\tilde{\beta}_i^*\}_{i=1}^I$ by applying S^* to the complete data set $\{X_i, Y_i\}_{i=1}^I$.

Confidence Intervals. Besides point estimates for individual-level partworths, our SL approach can also be used to produce confidence intervals for individual-level partworth estimates via bootstrapping, similar to the CO model (as detailed in the online appendix of Evgeniou et al. 2007). To generate the bootstrap estimates for confidence intervals, we first estimate the optimal modeling strategy $S^* = (Q^*, \psi^*, COV^*)$. Next, we generate a large number of (e.g., 1,000) random bootstrap samples from the original data set and apply the modeling strategy S^* to each bootstrap sample; here the bootstrap samples are obtained by keeping all respondents and for each respondent randomly sampling her conjoint profiles with replacement. We then use the empirical distributions of partworth estimates generated from the bootstrap samples to construct confidence intervals.

2.5. Summary

We briefly summarize our SL model Metric-SL in the following. The MATLAB code for Metric-SL is available from the authors on request.

First Stage.

Step 1a. Obtain the initial estimates $\{\bar{\beta}_i\}_{i=1}^I$ and the scaled covariance matrix of the partworths $\bar{D}$ using RR-Het (Evgeniou et al. 2007).

Step 1b. For each $\Gamma \in \Theta$, set $\theta_{ik} = e^{-\omega((\tilde{\beta}_i - \bar{\beta}_k)^\top \bar{D}^{-1}(\tilde{\beta}_i - \bar{\beta}_k))^{1/2}}$, and solve (Metric-SEG) (see (1)). Recover the candidate segmentation $Q(\Gamma)$ from $B^S(\Gamma)$.

Step 1c. Repeat Step 1b for each $\Gamma \in \Theta$, and obtain the set of candidate segmentations

$$\Phi = \{Q(\Gamma)\}_{\Gamma \in \Theta: Q(\Gamma) \neq \varnothing} \cup \{Q(\text{Trivial})\}. \tag{6}$$

Second Stage.

Step 2a. For each $Q \in \Phi$, define a set of modeling strategies $\{S \mid S = (Q, \psi, COV)$ s.t. $\psi^l \in \Xi$, $COV^l \in \{G, R\}$, for $l = 1, 2, \ldots, L\}$. A modeling strategy S recovers the individual-level partworths by solving a set of L optimization problems $\{\text{Metric-HET}(Q; l; \psi^l; COV^l)\}_{l=1}^{L}$ defined in (4) and (5).

Step 2b. Select the modeling strategy $S^* = (Q^*, \psi^*, COV^*)$ with the minimum cross-validation error, i.e., $S^* = \text{argmin}_S \text{CVE}(S)$. We select Q^* as the optimal segmentation.

Step 2c. Generate the optimal individual-level partworth estimates $\{\tilde{\beta}_i^*\}_{i=1}^{I}$ by then applying S^* to $\{X_i, Y_i\}_{i=1}^{I}$, i.e., by solving L^* optimization problems $\{\text{Metric-HET}(Q^*; l; \psi^{l*}; COV^{l*})\}_{l=1}^{L^*}$. The outputs of the second stage are $(\{\tilde{\beta}_i^*\}_{i=1}^{I}, Q^*)$, which are also the final outputs of the complete Metric-SL model.

2.6. Extension to Choice-Based Conjoint Analysis

CBC has been the dominant conjoint approach recently (Iyengar et al. 2008). Our SL model can be readily extended to the context of CBC. In particular, our SL model can be applied to CBC by simply replacing the squared-error loss functions in all optimization problems in Metric-SL with the logistic loss functions. We discuss our SL model for CBC, *Choice-SL*, in the Web appendix. The MATLAB code for Choice-SL is available from the authors on request.

3. Simulation Experiments

In this section we report the results of a set of simulation experiments designed to test the performance of our SL model. Simulation experiments have been widely adopted in the marketing literature to evaluate conjoint estimation methods (Vriens et al. 1996, Andrews et al. 2002b). We consider both metric and choice-based conjoint simulation experiments.

3.1. Metric Conjoint Simulation Experiments

We compared Metric-SL, the metric version of our SL model, to three benchmark methods: (1) the FM model (Kamakura and Russell 1989, Chintagunta et al. 1991), (2) the Bayesian NCM model (Allenby et al. 1998), and (3) RR-Het, the metric version of the CO model of Evgeniou et al. (2007). The FM model represents MCH using discrete mass points. The NCM model specifies a mixture of multivariate normal distributions to characterize the heterogeneity distribution and is capable of representing a wide variety of heterogeneity distributions. RR-Het is not specifically designed to model MCH; however, we included it as a benchmark method to assess the improvement made by adopting the more general Metric-SL model.

The implementation of the three benchmark methods closely followed the extant literature. In particular, the FM model was calibrated using the Bayesian information criterion (BIC) (Andrews et al. 2002b), and for the NCM model the number of components was selected using the deviance information criterion (DIC) (Spiegelhalter et al. 2002, Luo 2011). We provide the setup of the NCM model, including the specification of parameters for the second-stage priors, in the Web appendix.

3.1.1. Data. Our experimental design and data-generating process largely followed past work that has used simulations to evaluate methods for recovering MCH within metric conjoint settings (Andrews et al. 2002b). See Andrews et al. (2002b) for a discussion of the experimental design and the data-generating process.

Experimental design. We experimentally manipulated four data characteristics:

Factor 1. The number of segments: 2 or 3

Factor 2. The number of profiles per respondent (for calibration): 18 or 27

Factor 3. The error variance: 0.5 or 1.5

Factor 4. The within-segment variances of distributions: 0.05, 0.10, 0.20, 0.40, 0.60, 0.80, or 1.00

Hence, we used a $2^3 \times 7$ design, resulting in a total of 56 experimental conditions. We randomly generated 5 data sets for each experimental condition and estimated all conjoint models separately on each data set.

Data-generating process. We adopted the conjoint designs used in Andrews et al. (2002b) in which six product attributes were varied at three levels each. Each data set consisted of 100 synthetic respondents and their responses were generated according to the following three-step process: we (1) generated the true segment-level partworths, (2) assigned each respondent to a segment and generated her true individual-level partworths, and (3) generated her response vector. More specifically, the true segment-level partworths for any segment l, $\beta_l(S)$, were generated as a vector of random numbers sampled independently from a uniform distribution over the interval $[-1.7, 1.7]$. Each respondent was randomly assigned to all segments with equal probabilities, and her true individual-level partworths $\beta_i(T)$ were generated as $\beta_i(T) = \beta_l(S) + \sigma\xi_i$ if respondent i was assigned to segment l, where σ^2 is the prespecified within-segment variance (Factor 4) and ξ_i is a vector of independent standard normal random variables. Given $\beta_i(T)$, the response vector Y_i was computed as

$Y_i = X_i\beta_i(T) + \delta\epsilon_i$, where δ^2 is the prespecified error variance (Factor 3), and ϵ_i is a vector of independent standard normal random variables. To evaluate the predictive accuracy of the conjoint estimation methods, we generated 8 holdout profiles for each respondent regardless of whether 18 or 27 profiles (Factor 2) were used for calibration.

3.1.2. Results. We compared all four conjoint estimation methods in terms of parameter recovery and predictive accuracy. Parameter recovery was assessed using the root mean squared error (RMSE) between the true individual-level partworths $\beta_i(T)$ and the estimated individual-level partworths $\beta_i(E)$, which we denote by RMSE(β). Predictive accuracy was measured using the RMSE between the observed ratings $Y_i(O)$ and the predicted ratings $Y_i(P)$ on the holdout sample, which we denote by RMSE(Y). Following Evgeniou et al. (2007), we computed RMSE(β) and RMSE(Y) for each respondent in each data set and report the average RMSE(β) and RMSE(Y) across respondents and data sets for each experimental condition.[9]

Across experimental conditions, we find that Metric-SL overall outperforms the benchmark models both in terms of parameter recovery and predictive accuracy. In particular, Metric-SL performs best or not significantly different from best on RMSE(β) (at $p < 0.05$) in 51 out of 56 conditions, and is either the best performing method or indistinguishable from the best method on RMSE(Y) (at $p < 0.05$) in 52 out of 56 conditions. The comparisons are based on paired t-tests over the same 500 respondents, i.e., (100 respondents per data set) times (5 data sets), in each experimental condition.

To illustrate, we summarize the results for a subset of experimental conditions in Table 1, where *Num-S* denotes the number of segments in the heterogeneity distribution (Factor 1), *Num-P* denotes the number of profiles per respondent for calibration (Factor 2), *EV* denotes the error variance (Factor 3), and *WSV* denotes the within-segment variances of distributions (Factor 4). We note that for both RMSE(β) and RMSE(Y), lower numbers indicate better performance. The full results for all 56 conditions are reported in the Web appendix.

Table 1 shows a systematic pattern of RMSE(β) and RMSE(Y) for the four conjoint estimation methods with respect to *WSV*. When *WSV* is small, e.g., *WSV* = 0.05 or 0.10, the NCM model and RR-Het perform substantially worse than Metric-SL, whereas the FM model shows a good performance. As *WSV* increases, the relative performance of the NCM model and RR-Het gradually improves, and that of the FM model quickly deteriorates. On the other hand, Metric-SL demonstrates a consistently strong performance across the range of *WSV*. This performance pattern confirms the importance of explicitly modeling both across-segment and within-segment heterogeneity, and also endogenously selecting an adequate amount of shrinkage to recover individual-level partworths in modeling MCH. The FM model assumes a discrete heterogeneity distribution that does not allow for within-segment heterogeneity and hence is not capable of fully capturing the variations in consumer preferences when within-segment heterogeneity is substantial. RR-Het models consumer preferences using a UCH distribution, which does not accommodate across-segment heterogeneity and thus limits its performance when the underlying heterogeneity distribution is fairly discrete. The NCM model explicitly models both across-segment and within-segment heterogeneity, but is not capable of endogenously selecting the amount of shrinkage, since it is influenced by exogenously chosen parameters for the second-stage priors. In Table 1, the relatively inferior performance of the NCM model when within-segment heterogeneity is small or moderate suggests that the amount of shrinkage imposed by the NCM model is inadequate in these experimental conditions. This provides evidence that, consistent with findings in Evgeniou et al. (2007), the amount of shrinkage imposed by the NCM model can be inadequate depending on the characteristics of a conjoint data set. By contrast, our Metric-SL model addresses both modeling challenges and shows a robust performance across conditions.

Table 1. RMSE(β) and RMSE(Y) for a Subset of Experimental Conditions

				RMSE(β)				RMSE(Y)			
Num-S	*Num-P*	*EV*	*WSV*	Metric-SL	NCM	FM	RR-Het	Metric-SL	NCM	FM	RR-Het
2	18	1.5	0.05	**0.2315**	0.4074	0.2367	0.3459	**1.2656**	1.3637	1.2918	1.3377
			0.10	**0.2959**	0.4319	0.3233	0.3889	**1.2822**	1.3591	1.3423	1.3432
			0.20	**0.3706**	0.4654	0.4534	0.4492	**1.3544**	1.4062	1.5338	1.4087
			0.40	**0.4645**	0.4956	0.6218	0.4981	**1.4359**	1.4498	1.7215	1.4597
			0.60	**0.5039**	0.5175	0.7485	0.5254	**1.4108**	**1.4115**	1.8227	1.4202
			0.80	**0.5450**	**0.5446**	0.8737	0.5606	1.4659	**1.4584**	2.0126	1.4695
			1.00	**0.5644**	**0.5635**	0.9789	0.5712	**1.4590**	**1.4605**	2.1829	**1.4630**

Note. Bold numbers in each experimental condition for each performance measure indicate best or not significantly different from best at the $p < 0.05$ level based on paired t-tests.

We conducted a regression analysis to examine the impact of the experimental factors on RMSE(β) and RMSE(Y) of the four conjoint estimation methods. For RMSE(β), we adopted the following specification:

$$\begin{aligned}\text{RMSE}(\beta)_t &= \alpha_0 + \alpha_1 \times \text{Num-S-Dummy}_t \\ &\quad + \alpha_2 \times \text{Num-P-Dummy}_t + \alpha_3 \times \text{EV-Dummy}_t \\ &\quad + \alpha_4 \times \text{WSV}_t + \epsilon_t, \end{aligned} \tag{7}$$

where the index t runs over the 56 experimental conditions. The dependent variable RMSE(β)$_t$ is the average RMSE(β) of a method in condition t. For the independent variables, we dummy coded the first three experimental factors, *Num-S*, *Num-P*, and *EV*, and used the original value of the fourth experimental factor, *WSV*.[10] Table 2 shows the results of the ordinary least squares (OLS) estimation on RMSE(β) for each of the four conjoint estimation methods.

We make a few observations from the results in Table 2. The fact that the coefficients for *Num-S* are insignificant for all methods suggests that the number of segments has little impact on RMSE(β). *Num-P* has significant negative coefficients for all methods except the FM model, implying that more calibration profiles improve the accuracy of parameter recovery for the three methods other than the FM model. *EV* has significant positive coefficients for all methods, which means that a larger error variance hurts all methods; we note that the impact of error variance on the FM model is smaller compared to other methods. *WSV* has significant positive coefficients, indicating that a larger within-segment variance leads to a higher error in parameter recovery for all methods. Furthermore, as *WSV* increases, the FM model deteriorates most quickly, followed by Metric-SL, which is in turn followed by RR-Het and the NCM model. This is consistent with our previous findings about the relative performance of the four conjoint estimation methods with respect to *WSV*.

We also conducted a regression analysis to understand the impact of the experimental factors on the relative performance between Metric-SL and the NCM model. In particular, we adopted a specification identical to (7) except that the dependent variable was replaced with the difference of RMSE(β) for Metric-SL and the NCM model. The results of the OLS estimation are reported in the last column of Table 2. The results show that the performance of Metric-SL relative to the NCM model improves when there are fewer calibration profiles. This finding highlights the usefulness of Metric-SL especially in contexts where researchers prefer to elicit consumer preferences using short conjoint questionnaires due to concerns over response rates and response quality (Lenk et al. 1996). We also find that a larger error variance and a smaller within-segment variance improve the relative performance of Metric-SL.

For RMSE(Y), we used a specification identical to (7) except that the dependent variable was the average RMSE(Y) of a method in a specified experimental condition. We report the results of the OLS estimation in Table 3.

The impact of the experimental factors on RMSE(Y) is largely similar to that on RMSE(β). A couple of main differences are that for RMSE(Y), *Num-S* has significant positive coefficients for all methods except Metric-SL, and *Num-P* has the largest impact on the FM model.

3.2. Choice-Based Conjoint Simulation Experiments

We compared Choice-SL, the choice version of our SL model, to three benchmark methods: (1) the FM model, (2) the NCM model, and (3) LOG-Het, the choice version of the CO model. All benchmark methods were the choice versions of those in Section 3.1, and the implementations were similar to their metric version counterparts.

3.2.1. Data. Our experimental design and data-generating process largely followed past work that used simulations to evaluate methods for recovering MCH

Table 2. Regression Analysis of RMSE(β) for Metric Simulations

Variable	Metric-SL	NCM	FM	RR-Het	Metric-SL – NCM
Intercept	0.2033***	0.2884***	0.2392***	0.2582***	−0.0851***
Num-S	0.0062	0.0068	0.0146	0.0105	−0.0006
Num-P	−0.0427***	−0.0609***	−0.0123	−0.0589***	0.0181**
EV	0.1217***	0.1493***	0.0253**	0.1471***	−0.0277***
WSV	0.2254***	0.1106***	0.7622***	0.1539***	0.1147***
R^2	0.86	0.97	0.98	0.93	0.69

Notes. The dependent variables in the second, third, fourth, and fifth columns are RMSE(β)'s of Metric-SL, NCM, FM, and RR-Het, respectively; the dependent variable in the sixth column is the difference between RMSE(β)'s of Metric-SL and NCM.
**$p < 0.05$; ***$p < 0.01$.

Table 3. Regression Analysis of RMSE(Y) for Metric Simulations

Variable	Metric-SL	NCM	FM	RR-Het	Metric-SL − NCM
Intercept	0.7662***	0.8154***	0.8861***	0.8024***	−0.0492***
Num-S	0.0108	0.0119**	0.0327**	0.0140**	−0.0012
Num-P	−0.0488***	−0.0607***	−0.1058***	−0.0632***	0.0120**
EV	0.5497***	0.5636***	0.3989***	0.5639***	−0.0139***
WSV	0.1395***	0.0728***	0.9374***	0.0954***	0.0667***
R^2	0.99	0.99	0.98	0.99	0.67

Notes. The dependent variables in the second, third, fourth, and fifth columns are RMSE(Y)'s of Metric-SL, NCM, FM, and RR-Het, respectively; the dependent variable in the sixth column is the difference between RMSE(Y)'s of Metric-SL and NCM.
$^{**}p < 0.05$; $^{***}p < 0.01$.

using choice data (Andrews et al. 2002a, Andrews and Currim 2003).

Experimental design. We experimentally manipulated four data characteristics:

Factor 1. The number of segments: 2 or 3

Factor 2. The number of choice sets per respondent (for calibration): 16 or 24

Factor 3. The error variance: standard (1.645) or high (3.290)

Factor 4. The within-segment variances of distributions: 0.05, 0.10, 0.20, 0.40, 0.60, 0.80, or 1.00

Hence, we used a $2^3 \times 7$ design, resulting in a total of 56 experimental conditions. We randomly generated 5 data sets for each experimental condition and estimated all conjoint models separately on each data set.

Data-generating process. In all data sets, each choice set consisted of four conjoint profiles, each associated with a distinct brand. In addition to the three (i.e., $4-1=3$) brand dummies, the attributes also included one continuous variable and two binary variables. We created four levels for the continuous variable, each being a range: "low" $\equiv [-1.3, -0.65]$, "medium-low" $\equiv [-0.65, 0]$, "medium-high" $\equiv [0, 0.65]$, and "high" $\equiv [0.65, 1.3]$. For each choice set, we randomly selected a value from each range and assigned the four values to the profiles such that each profile had an equal chance to be assigned with the lowest value. For each of the two binary attributes, we randomly selected a profile in a choice set and set its value on the attribute to 1. We note that the design of the continuous attribute and the two binary attributes was aimed at inducing sufficient variations in the data and is different from those in Andrews et al. (2002a) and Andrews and Currim (2003), which consider scanner panel applications rather than conjoint applications.

In each data set, the choices of 100 synthetic respondents were generated using a three-step process similar to that in Section 3.1. We closely followed Andrews et al. (2002a) and Andrews and Currim (2003) and generated three levels of segment-level coefficients (low, medium, and high) for each of the six attributes (i.e., three brand dummies, one continuous variable, and two binary variables). The rationale for this design with three levels of coefficients was to have different segments assigned with distinct levels of coefficients for each attribute and therefore create clear separations between segments. The medium-level coefficients were generated as follows: the brand-specific constants were sampled from a uniform distribution over the interval $[-1, 1]$, the coefficient of the continuous variable was sampled from a uniform distribution over $[-2.5, -2]$, and the coefficients of the binary variables were sampled from a uniform distribution over $[2, 2.5]$. The high-level (respectively, low-level) coefficients were generated by adding to (respectively, subtracting from) the corresponding medium-level coefficients a normal random variable drawn from $N(1.5, 0.15^2)$, where 1.5 was the mean separation between segments (Andrews et al. 2002a, Andrews and Currim 2003). In experimental conditions with three segments (Factor 1), we generated the true segment-level partworths by assigning the three levels of coefficients of each attribute randomly to the three segments. In experimental conditions with two segments, we simply retained the true segment-level partworths of the first two segments generated in the three-segment conditions. We denote the true segment-level partworths for any segment l as $\beta_l(S)$.

Each respondent was randomly assigned to the available segments with equal probabilities. As in Section 3.1, respondent i's true individual-level partworths $\beta_i(T)$ were generated as $\beta_i(T) = \beta_l(S) + \sigma\xi_i$ if respondent i was assigned to segment l, where σ^2 is the prespecified within-segment variance (Factor 4) and ξ_i is a vector of independent standard normal random variables. Given $\beta_i(T)$, respondent i's choices were stochastically generated according to the logit model where the variance of the type-I extreme value random variables was given by the prespecified error variance (Factor 3). To evaluate the predictive accuracy of the conjoint estimation methods, we generated 8 holdout choice sets for each respondent regardless of whether 16 or 24 choice sets (Factor 2) were used for calibration.

3.2.2. Results. We compared the four conjoint estimation methods in terms of parameter recovery and

Table 4. RMSE(β) and Holdout-LL for a Subset of Experimental Conditions

				RMSE(β)				Holdout-LL			
Num-S	*Num-CS*	*EV*	*WSV*	Choice-SL	NCM	FM	LOG-Het	Choice-SL	NCM	FM	LOG-Het
2	16	3.290	0.05	0.5521	0.6760	**0.3598**	0.7021	−0.8900	−0.9207	**−0.8765**	−0.9342
			0.10	0.4920	0.6692	**0.3584**	0.6429	−0.9069	−0.9397	**−0.8972**	−0.9427
			0.20	0.5087	0.6829	**0.4646**	0.6738	**−0.9131**	−0.9420	**−0.9165**	−0.9410
			0.40	**0.6389**	0.7773	0.6602	0.7418	**−0.8758**	−0.8966	−0.9092	−0.8876
			0.60	**0.7315**	0.8512	0.8324	0.7820	**−0.9109**	−0.9351	−0.9526	−0.9199
			0.80	**0.7653**	0.8053	0.9207	0.8068	**−0.9229**	−0.9276	−0.9980	**−0.9204**
			1.00	**0.8783**	**0.8791**	0.9990	0.9045	**−0.9237**	−0.9312	−0.9808	**−0.9219**

Notes. Bold numbers in each experimental condition for each performance measure indicate best or not significantly different from best at the $p < 0.05$ level based on paired t-tests.

predictive accuracy. Parameter recovery was assessed using RMSE(β). Predictive accuracy was measured using the holdout sample log-likelihood (Andrews et al. 2002a), which we denote by *Holdout-LL*. Again, for each experimental condition, we report the average RMSE(β) and Holdout-LL across all respondents and data sets.[11]

Similar to metric simulation experiments, we find that Choice-SL overall outperforms the benchmark models both in terms of parameter recovery and predictive accuracy. In particular, Choice-SL is the best performing model or indistinguishable from the best model on RMSE(β) (at $p < 0.05$) in 31 out of 56 conditions, and performs best or not significantly different from best on Holdout-LL (at $p < 0.05$) in 42 out of 56 conditions.

For the purpose of illustration, we summarize the results for a subset of experimental conditions in Table 4, where *Num-S* denotes the number of segments in the heterogeneity distribution (Factor 1), *Num-CS* denotes the number of choice sets per respondent for calibration (Factor 2), *EV* denotes the error variance (Factor 3), and *WSV* denotes the within-segment variances of distributions (Factor 4). We note that for RMSE(β), lower numbers indicate better performance, whereas for Holdout-LL, higher numbers indicate better performance. The full results for all 56 conditions are reported in the Web appendix.

We find that results in Table 4 are qualitatively similar to those in Table 1 except that the FM model becomes the best performing model when *WSV* is small.

As in the metric simulation experiments, we conducted a regression analysis to examine the impact of the experimental factors on both performance measures, RMSE(β) and Holdout-LL. The regression specifications were similar to (7). Tables 5 and 6 report the results of the OLS estimation.

Table 5 shows that the performance of Choice-SL relative to the NCM model in terms of parameter recovery improves with more segments and a smaller within-segment variance. Table 6 shows that the performance of Choice-SL relative to the NCM model in terms of predictive accuracy improves with fewer choice sets for calibration. This finding, consistent with what we found in the metric simulation experiments, further emphasizes the usefulness of our model in contexts in which concerns over response rates and response quality prompt researchers to use short conjoint questionnaires. We also find that more segments and a smaller within-segment variance improve the relative performance of Choice-SL. In Section 4, we leverage these findings to explain the relative performance among models on field data.

4. Field Data

4.1. Metric Conjoint

We evaluate the performance of our Metric-SL model using a metric conjoint data set of personal computers

Table 5. Regression Analysis of RMSE(β) for Choice-Based Simulations

Variable	Choice-SL	NCM	FM	LOG-Het	Choice-SL − NCM
Intercept	0.4446***	0.6287***	0.3143***	0.6153***	−0.1841***
Num-S	0.0270**	0.0668***	0.0364**	0.0535***	−0.0398**
Num-CS	−0.0919***	−0.1093***	−0.0450***	−0.0958***	0.0174
EV	0.0426***	0.0658***	0.0070	0.0415***	−0.0232
WSV	0.3834***	0.1405***	0.7626***	0.2230***	0.2429***
R^2	0.94	0.88	0.96	0.93	0.71

Notes. The dependent variables in the second, third, fourth, and fifth columns are RMSE(β)'s of Choice-SL, NCM, FM, and LOG-Het, respectively; the dependent variable in the sixth column is the difference between RMSE(β)'s of Choice-SL and NCM.

**$p < 0.05$; ***$p < 0.01$.

Table 6. Regression Analysis of Holdout-LL for Choice-Based Simulations

Variable	Choice-SL	NCM	FM	LOG-Het	Choice-SL − NCM
Intercept	−0.7234***	−0.7580***	−0.7156***	−0.7609***	0.0346***
Num-S	−0.0002	−0.0067	−0.0034	−0.0035	0.0064***
Num-CS	0.0117	0.0175**	0.0059	0.0164**	−0.0058**
EV	−0.1888***	−0.1899***	−0.1713***	−0.1874***	0.0011
WSV	0.0058	0.0370***	−0.1037***	0.0473***	−0.0312***
R^2	0.91	0.92	0.88	0.92	0.68

Notes. The dependent variables in the second, third, fourth, and fifth columns are Holdout-LL's of Choice-SL, NCM, FM, and LOG-Het, respectively; the dependent variable in the sixth column is the difference between Holdout-LL's of Choice-SL and NCM.
**$p < 0.05$; ***$p < 0.01$.

that was first introduced in Lenk et al. (1996). The same data set was also used in Evgeniou et al. (2007) to compare conjoint estimation methods. In the study, 180 respondents each rated 20 hypothetical personal computers on an 11-point scale (0 to 10). Each hypothetical profile was represented using 13 binary attributes and an intercept. The first 16 profiles formed an orthogonal and balanced design and were used for calibration, and the last 4 were used for holdout validation. See Lenk et al. (1996) and Evgeniou et al. (2007) for details of this data set.

We compared the predictive accuracy of four models, Metric-SL, the FM model, the NCM model, and RR-Het using RMSE(Y) and the first choice hits in the holdout sample (Andrews et al. 2002b), which we denote by *1stCH*. For any respondent, 1stCH was set to 1 if the holdout profile with the highest observed rating was correctly predicted and 0 otherwise. We report the average RMSE(Y) and 1stCH across 180 respondents for each method. Table 7 summarizes the results. We note that for RMSE(Y), lower numbers indicate better performance, whereas for 1stCH, higher numbers indicate better performance.

Using paired t-tests over the 180 respondents, we find that Metric-SL and RR-Het perform best or not significantly different from best (at $p < 0.10$) both in terms of RMSE(Y) and 1stCH. This performance comparison validates the predictive accuracy of Metric-SL; it also suggests that the assumption of a UCH distribution made by RR-Het is not restrictive on this data set.

Table 7. Field Conjoint Data Sets

The personal computer data set				
	Metric-SL	NCM	FM	RR-Het
RMSE(Y)	**1.6099**	1.6558***	1.8639***	**1.6072**
1stCH	**0.7056**	0.6722**	0.5889***	**0.6944**
The hotel data set				
	Choice-SL	NCM	FM	LOG-Het
Holdout-LL	**−0.9270**	−1.0297**	**−0.9192**	**−0.9305**
Holdout-HIT	**0.6330**	**0.6410**	0.5878**	0.6090*
The cell phone plan data set				
	Choice-SL	NCM	FM	LOG-Het
Holdout-LL	**−0.9205**	−0.9540*	−0.9944***	−0.9389*
Holdout-HIT	0.6278*	**0.6407**	0.5856***	0.6190***

Notes. For RMSE(Y), lower numbers indicate better performance; for 1stCH, higher numbers indicate better performance; for Holdout-LL, higher numbers indicate better performance; and for Holdout-HIT, higher numbers indicate better performance. Bold numbers indicate best or not significantly different from best at the $p < 0.10$ level.
*Significantly different from best at the $p < 0.10$ level; **significantly different from best at the $p < 0.05$ level; ***significantly different from best at the $p < 0.01$ level.

4.2. Choice-Based Conjoint

4.2.1. Application 1—Hotel Choice. A total of 188 respondents participated in this study, and each of them was shown 12 choice sets. Each choice set consisted of three hotel profiles and a no-choice option. Seven attributes, including brand, room rate, location, restaurant, gym, Internet access, and rewards points, were used to represent the profiles. The brand attribute was treated as discrete with five levels, e.g., Westin, whereas all other attributes were treated as continuous. We randomly selected 10 out of the 12 choice sets for each respondent for calibration and used the remaining 2 choice sets for holdout validation.

We compared the predictive performance of four models—Choice-SL, the FM model, the NCM model, and LOG-Het—using Holdout-LL and the holdout sample hit rate, which we denote by *Holdout-HIT*. We report the average Holdout-LL and Holdout-HIT across all 188 respondents for each method. The results are summarized in Table 7. We note that for both Holdout-LL and Holdout-HIT, higher numbers indicate better performance. Using paired t-tests over 188 respondents, we find that Choice-SL performs best or not significantly different from best (at $p < 0.10$) for both Holdout-LL and Holdout-HIT. The NCM model performs significantly worse than best on Holdout-LL, and the FM model and LOG-Het perform significantly worse than best on Holdout-HIT. Thus, the empirical

performance comparison validates the predictive accuracy of Choice-SL.

4.2.2. Application 2—Cell Phone Plan Choice. A total of 72 respondents participated in this study, and each of them was shown 18 choice sets that consisted of three profiles and a no-choice option. Six attributes were used for constructing the conjoint profiles: access fee, per-minute rate, plan minutes, service provider, Internet access, and rollover of unused minutes. The same data set was used in Iyengar et al. (2008). We used the best fitting "nonlinear-effects" specification in Iyengar et al. (2008) that adds logarithmic terms in access fee, per-minute rate, and plan minutes to the standard conjoint specification.[12] We randomly selected 15 out of the 18 choice sets for each respondent for calibration and used the remaining 3 choice sets for holdout validation.

We compared the predictive performance of four models—Choice-SL, the FM model, the NCM model, and LOG-Het—using Holdout-LL and Holdout-HIT. Since the sample size of this data set is relatively small, the paired t-tests among the four models were insignificant on both performance measures. To tackle this issue, we adopted the following alternative statistical test procedure. We generated 10 random replications of the data set. In each replication, we retained all 72 respondents, randomly selected 15 out of the 18 choice sets for each respondent for calibration, and used the remaining 3 choice sets for holdout validation. Each conjoint estimation method was separately applied to each of the 10 replications, and Holdout-LL and Holdout-HIT for each respondent were computed in each replication. We computed the average Holdout-LL and Holdout-HIT across 10 replications for each respondent and compared the four conjoint estimation methods using paired t-tests over the 72 respondents. The results are summarized in Table 7. Recall that for both Holdout-LL and Holdout-HIT, higher numbers indicate better performance. We see from Table 7 that Choice-SL performs best (at $p < 0.10$) on Holdout-LL and the NCM model performs best (at $p < 0.10$) on Holdout-HIT.

4.2.3. Comparison Between the Two Choice-Based Applications. Table 7 shows that the predictive accuracy of Choice-SL compared to the NCM model is more favorable on the hotel data set than on the cell phone plan data set. It is instructive to interpret this comparison using our findings in Section 3.2 regarding how the predictive performance of Choice-SL relative to the NCM model varies with respect to the data characteristics. First, Choice-SL recovers 2 segments in the hotel data set as well as in most replications of the cell phone plan data set, and hence there is no clear evidence suggesting that the two data sets have different numbers of segments. Second, the number of calibration choice sets of the hotel data set (i.e., 10) is smaller than that of the cell phone plan data set (i.e., 15). Third, we use Choice-SL to infer the within-segment variances in both data sets and find that the average inferred within-segment variance for the hotel data set is smaller than that for the cell phone plan data set. Recall that in Section 3.2 we found that Choice-SL is likely to perform better relative to the NCM model in terms of predictive accuracy when the number of calibration choice sets is small and the within-segment variance is small. Therefore, the results in Table 7 are consistent with our findings in the simulation experiments.

4.3. Graphical Illustration of Partworth Estimates

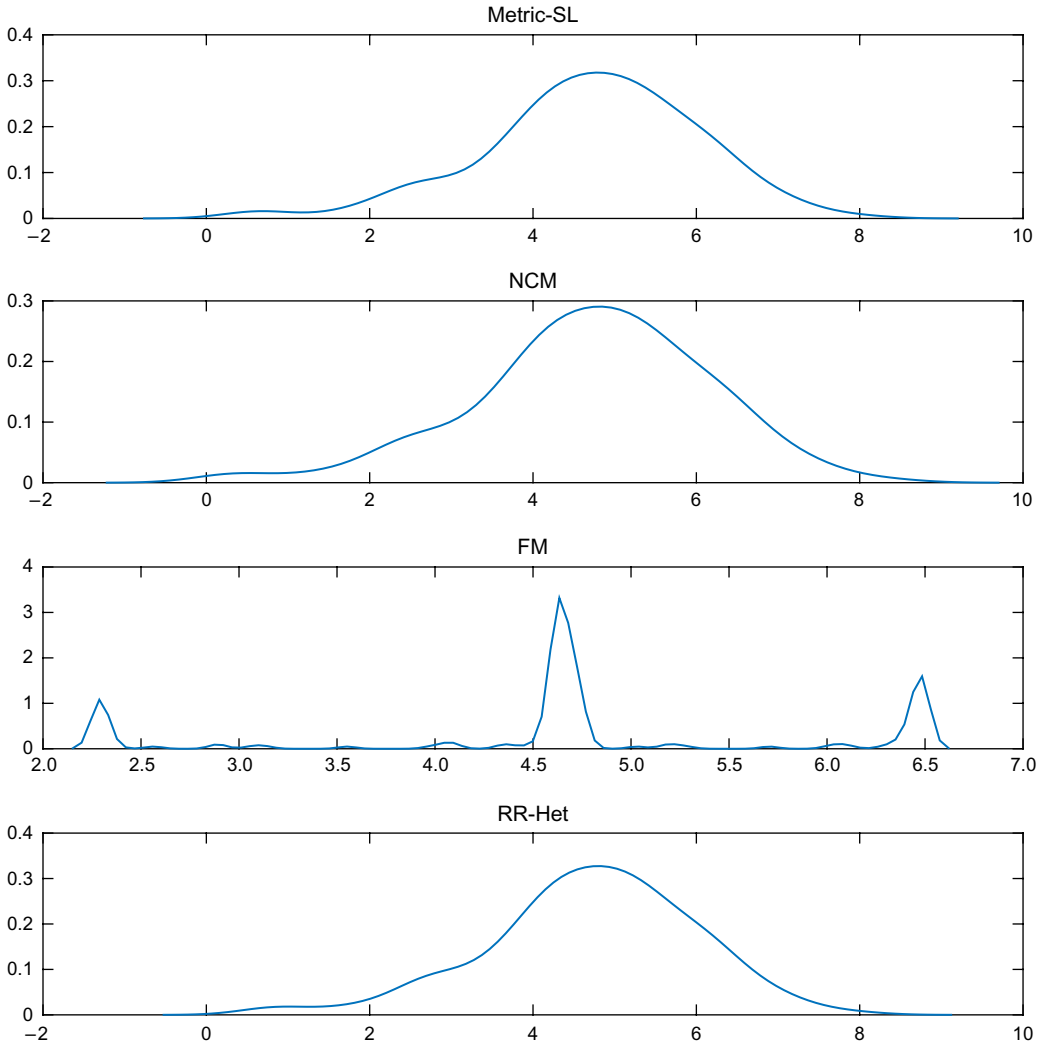
In this section, we provide graphical illustrations of the individual-level heterogeneity representations recovered by the four methods on the three field data sets. Given a conjoint estimation method and a data set, we estimate a density for each partworth by applying a kernel smoothing density estimator to the individual-level point estimates of the partworth for all respondents.

To illustrate, we plot the density estimates for the following partworths. Figure 1 displays the density of intercept in the personal computer data set. The density curves estimated by Metric-SL, the NCM model, and RR-Het are qualitatively similar and exhibit largely unimodal continuous shapes, while the FM model recovers three spikes in the density curve. Figure 2 shows the density of the partworth corresponding to location in the hotel data set. It is evident that the density curves estimated by Choice-SL and the FM model display multimodal continuous shapes, whereas those estimated by the NCM model and LOG-Het are unimodal. In Figure 3, we plot the density of the partworth corresponding to plan minutes in the cell phone plan data set. The density curves of all methods except LOG-Het show multimodal continuous shapes, with the multimodality estimated by Choice-SL and the FM model being more pronounced than that estimated by the NCM model. Density estimates for other partworths are available from the authors on request.

4.4. Comparison on Pricing Implications

We use the hotel data set as an example to compare the pricing implications of the four conjoint estimation methods. To illustrate, we consider a hotel profile with the following attributes: brand set to Westin, location and Internet access set to high levels, and restaurant, gym, and rewards points set to medium levels. We use the individual-level partworth estimates obtained from each method to derive the individual-level willingness to pay (WTP) defined as the price at which a respondent is indifferent between choosing

Figure 1. (Color online) Density Plots: Intercept in the Personal Computer Data Set



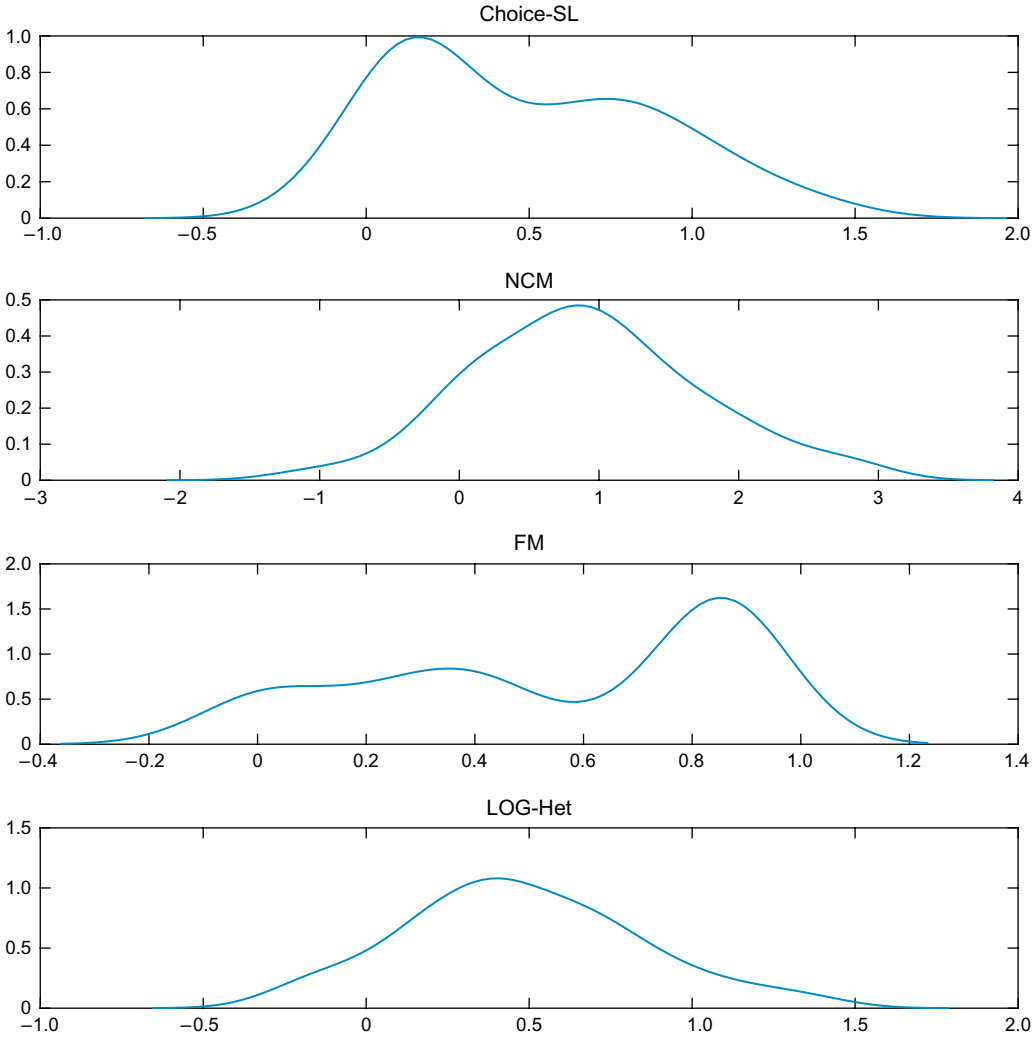
this particular hotel profile and the no-choice option (Jedidi and Zhang 2002). To ensure that the WTP estimates are plausible, we set the minimum (respectively, maximum) feasible WTP to \$0 (respectively, \$1,000).

We find that the primary distinguishing characteristic between the WTPs estimated by Choice-SL and the other three methods is that the latter infer a large WTP for more respondents. Choice-SL infers that 34.6% of respondents have a WTP greater than \$300, whereas the NCM model, the FM model, and LOG-Het estimate this proportion to be 50.5%, 41.0%, and 43.1%, respectively. This difference in the estimates for the fraction of respondents with large WTPs has a substantial impact on the revenue-maximizing prices implied by different methods.[13] Choice-SL, the NCM model, the FM model, and LOG-Het set the revenue-maximizing prices to \$216, \$477, \$458, and \$790, respectively. Furthermore, Choice-SL, the NCM model, the FM model, and LOG-Het estimate the proportion of respondents who would prefer the hotel profile described above to the no-choice option at the revenue-maximizing prices to be 55.3%, 33.5%, 37.2%, and 17.6%, respectively. Hence, the four conjoint estimation methods imply different pricing strategies. Choice-SL recommends using a moderate price to capture a large chunk of the market, whereas the other three methods (especially LOG-Het) recommend using a high price to extract revenue from a smaller segment of respondents with high WTPs. Given that the highest price shown in all hotel profiles was \$250, we find that the pricing decision of Choice-SL has higher face validity.

5. Conclusions

Consumer preferences can often be modeled using an MCH distribution, and adequate modeling of MCH is critical for accurate conjoint estimation. In this paper, we propose an innovative SL approach for modeling MCH. The SL approach models MCH via a two-stage divide-and-conquer framework, in which

Figure 2. (Color online) Density Plots: The Partworth Corresponding to Location in the Hotel Data Set

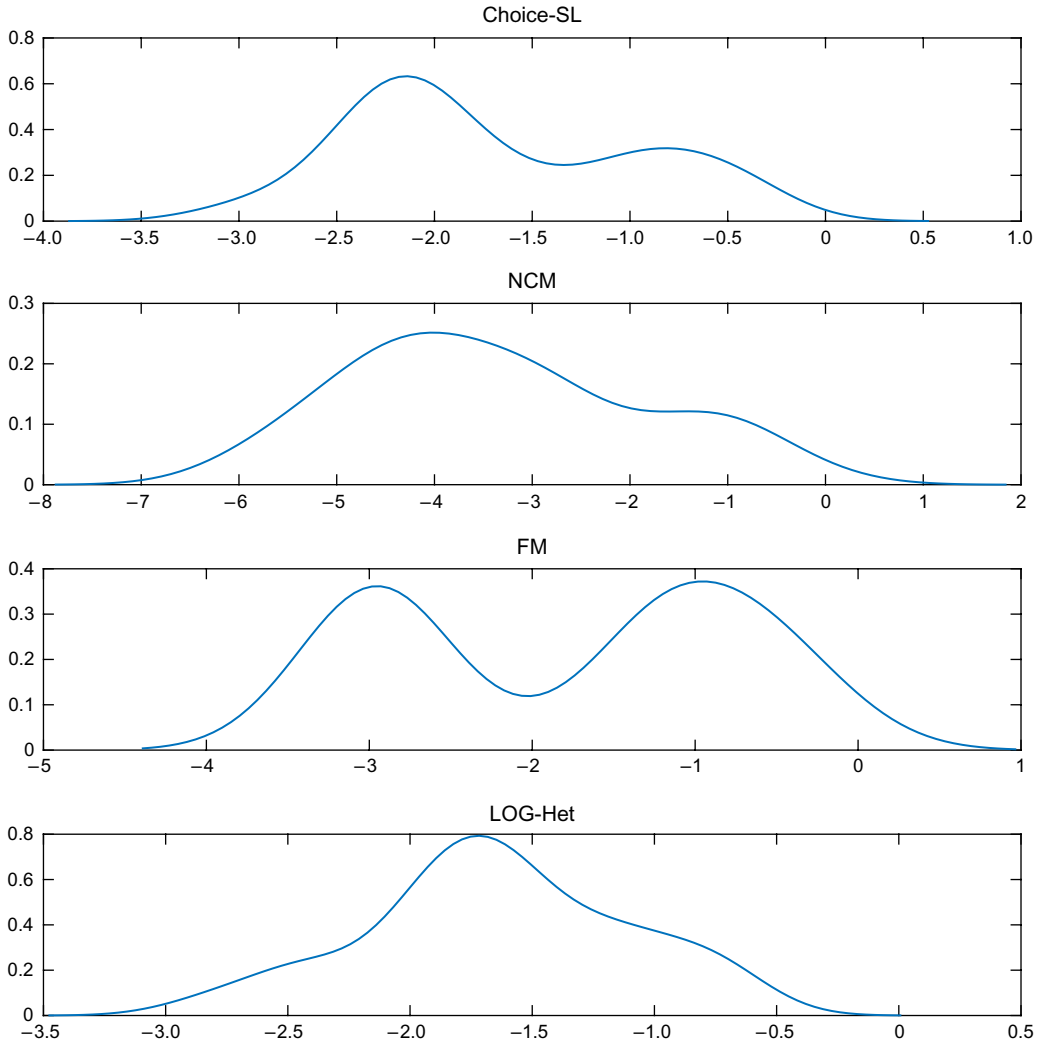


MCH is decomposed into a small collection of within-segment UCH distributions using sparse learning methodology, and each UCH is then modeled separately. Consequently, we explicitly account for both across-segment and within-segment heterogeneity in the SL model. In addition, the amount of shrinkage imposed to recover the individual-level partworths is endogenously selected using cross-validation.

We test the empirical performance of our SL model and compare it to the finite mixture model (Kamakura and Russell 1989, Chintagunta et al. 1991), the Bayesian normal component mixture model (Allenby et al. 1998), and the convex optimization model of Evgeniou et al. (2007) using extensive simulation experiments and three field data sets. We find that our SL model demonstrates a consistently strong performance across a wide range of experimental conditions as well as field data sets with distinct characteristics. We also show the managerial relevance of our SL model using an optimal pricing exercise in which the SL model generates a more plausible revenue-maximizing price.

There are several promising avenues for future research. First, we can consider an extension of our SL model by incorporating kernel methods (Vapnik 1998), which were introduced to marketing by Cui and Curry (2005) and Evgeniou et al. (2005). Second, researchers can also consider other population-based complexity controls to improve the capability for modeling MCH. Third, our SL model, like the finite mixture model and the Bayesian normal component mixture model, can be applied to estimate consumers' heterogeneous preferences in settings other than conjoint analysis, e.g., scanner panel data sets, and it may be fruitful to compare our SL model with benchmark models in such settings. Finally, an interesting research direction is to explore the potential of machine learning methods in modeling other phenomena in marketing beyond consumer heterogeneity.

Figure 3. (Color online) Density Plots: The Partworth Corresponding to Plan Minutes in the Cell Phone Plan Data Set



Acknowledgments

The authors thank Rick Andrews for sharing the conjoint designs used in Section 3.1, Peter Lenk for sharing the personal computer data set used in Section 4.1, and Rajan Sambandam from TRC Market Research for sharing the hotel data set used in Section 4.2.1. The authors also thank Eric Bradlow, Jeff Cai, Daria Dzyabura, Arun Gopalakrishnan, and Olivier Toubia for their insightful comments. The first two authors acknowledge the generous support of the Alex Panos Research Fund of the Wharton School of the University of Pennsylvania.

Endnotes

[1] Applications of nonparametric Bayesian methods in marketing include the Dirichlet process mixture model (Ansari and Mela 2003, Kim et al. 2004) and the centered Dirichlet process mixture model (Li and Ansari 2013). While nonparametric Bayesian methods provide more flexibility, they still suffer from the same limitations faced by the NCM model. With ongoing research in this area, we expect to see systematic comparisons between the benefits of using parametric and nonparametric Bayesian methods. In this paper, we compare our model with the FM and NCM models, which are more established modeling frameworks.

[2] We discuss the rationale behind sparsity-inducing penalty functions in the Web appendix.

[3] The specification for $\{\theta_{ik}\}$ in (3) uses only information contained in the conjoint data. Other information sources, e.g., consumers' demographic variables, can be readily incorporated in the specification for $\{\theta_{ik}\}$, and hence our SL model via a simple extension of (3). We discuss the extension in the Web appendix.

[4] We note that in (Metric-SEG), the amount of penalty imposed on $\|\beta_i^S - \beta_k^S\|_2$ is controlled by $\lambda\theta_{ik}$. In the empirical implementation of our SL model, we normalize $\theta = (\theta_{ik})$ such that $||\theta||_2 = 1$ and interpret the regularization parameter λ as controlling the "total" amount of penalty imposed on $\|\beta_i^S - \beta_k^S\|_2$'s.

[5] The specification for Θ used in simulation experiments and field applications is summarized in the Web appendix.

[6] In the empirical implementation of our SL model, we set $M = 10\%I$, such that any valid segment contains a nonnegligible portion of the population. The simulation experiments and field applications confirm the effectiveness of our choice of M.

[7] Recall that we also obtain a set of individual-level partworth estimates $\{B(\Gamma)\}$ by solving (Metric-SEG). We retain only the set of candidate segmentations Φ and exclude $\{B(\Gamma)\}$ as the output of the first stage because the latter are biased. We provide a detailed discussion about the bias in $\{B(\Gamma)\}$ in the Web appendix.

[8] The specification for Ξ used in simulation experiments and field applications is summarized in the Web appendix.

[9] In addition to parameter recovery and predictive accuracy, we also compared the computation times of Metric-SL and the NCM model and report the results in the Web appendix.

[10] $Num\text{-}S\text{-}Dummy_t = 1$ when $Num\text{-}S_t = 3$; $Num\text{-}P\text{-}Dummy_t = 1$ when $Num\text{-}P_t = 27$; and $EV\text{-}Dummy_t = 1$ when $EV_t = 1.5$.

[11] In addition to parameter recovery and predictive accuracy, we also compared the computation time of Choice-SL and the NCM model and report the results in the Web appendix.

[12] We differed from Iyengar et al. (2008) in that we standardized all continuous attributes, i.e., each continuous attribute was demeaned and divided by its standard deviation, before model estimation. The standardization is a widely adopted technique in the statistics and machine learning literature (Tibshirani 1996) that ensures that all continuous attributes have similar scales.

[13] If cost data were present, we could determine the profit-maximizing price.

References

Allenby GM, Rossi PE (1998) Marketing models of consumer heterogeneity. *J. Econometrics* 89(1):57–78.

Allenby GM, Arora N, Ginter JL (1998) On the heterogeneity of demand. *J. Marketing Res.* 35(3):384–389.

Andrews RL, Currim IS (2003) A comparison of segment retention criteria for finite mixture logit models. *J. Marketing Res.* 40(2):235–243.

Andrews RL, Ainslie A, Currim IS (2002a) An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity. *J. Marketing Res.* 39(4): 479–487.

Andrews RL, Ansari A, Currim IS (2002b) Hierarchical Bayes versus finite mixture conjoint analysis models: A comparison of fit, prediction, and partworth recovery. *J. Marketing Res.* 39(1): 87–98.

Ansari A, Mela CF (2003) E-customization. *J. Marketing Res.* 40(2): 131–145.

Argyriou A, Evgeniou T, Pontil M (2008) Convex multi-task feature learning. *Machine Learn.* 73(3):243–272.

Bach F, Jenatton R, Mairal J, Obozinski G (2011) Convex optimization with sparsity-inducing norms. Sra S, Nowozin S, Wright SJ, eds. *Optimization for Machine Learning* (MIT Press, Cambridge MA), 19–53.

Boyd S, Vandenberghe L (2004) *Convex Optimization* (Cambridge University Press, New York).

Celeux G, Hurn M, Robert CP (2000) Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* 95(451):957–970.

Chen X, Lin Q, Kim S, Carbonell JG, Xing EP (2012) Smoothing proximal gradient method for general structured sparse regression. *Ann. Appl. Statist.* 6(2):719–752.

Chintagunta PK, Jain DC, Vilcassim NJ (1991) Investigating heterogeneity in brand preferences in logit models for panel data. *J. Marketing Res.* 28(4):417–428.

Cui D, Curry D (2005) Prediction in marketing using the support vector machine. *Marketing Sci.* 24(4):595–615.

Evgeniou T, Boussios C, Zacharia G (2005) Generalized robust conjoint estimation. *Marketing Sci.* 24(3):415–429.

Evgeniou T, Pontil M, Toubia O (2007) A convex optimization approach to modeling consumer heterogeneity in conjoint estimation. *Marketing Sci.* 26(6):805–818.

Green PE, Srinivasan V (1990) Conjoint analysis in marketing: New developments with implications for research and practice. *J. Marketing* 54(4):3–19.

Hastie T, Tibshirani R, Friedman J (2001) *The Elements of Statistical Learning*, Springer Series in Statistics (Springer-Verlag, New York).

Hocking TD, Joulin A, Bach F, Vert JP (2011) Clusterpath: An algorithm for clustering using convex fusion penalties. *Proc. 28th Internat. Conf. Machine Learn., Bellevue, WA*.

Iyengar R, Jedidi K, Kohli R (2008) A conjoint approach to multipart pricing. *J. Marketing Res.* 45(2):195–210.

Jedidi K, Zhang ZJ (2002) Augmenting conjoint analysis to estimate consumer reservation price. *Management Sci.* 48(10):1350–1368.

Jenatton R, Gramfort A, Michel V, Obozinski G, Eger E, Bach F, Thirion B (2012) Multiscale mining of fMRI data with hierarchical structured sparsity. *SIAM J. Imaging Sci.* 5(3):835–856.

Kamakura WA, Russell GJ (1989) A probabilistic choice model for market segmentation and elasticity structure. *J. Marketing Res.* 26(4):379–390.

Kim JG, Menzefricke U, Feinberg FM (2004) Assessing heterogeneity in discrete choice models using a Dirichlet process prior. *Rev. Marketing Sci.* 2(1):1–39.

Kim S, Xing EP (2012) Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping. *Ann. Appl. Statist.* 6(3):1095–1117.

Lenk PJ, DeSarbo WS, Green PE, Young MR (1996) Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Sci.* 15(2): 173–191.

Li Y, Ansari A (2013) A Bayesian semiparametric approach for endogeneity and heterogeneity in choice models. *Management Sci.* 60(5):1161–1179.

Luo L (2011) Product line design for consumer durables: An integrated marketing and engineering approach. *J. Marketing Res.* 48(1):128–139.

Qin Z, Goldfarb D (2012) Structured sparsity via alternating direction methods. *J. Machine Learn. Res.* 13(1):1435–1468.

Rossi PE, Allenby GM, McCulloch R (2005) *Bayesian Statistics and Marketing* (John Wiley & Sons, West Sussex, UK).

Rossi PE, McCulloch RE, Allenby GM (1996) The value of purchase history data in target marketing. *Marketing Sci.* 15(4):321–340.

Shao J (1993) Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* 88(422):486–494.

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. *J. Roy. Statist. Soc.: Ser. B* (*Statist. Methodology*) 64(4):583–639.

Stephens M (2000) Dealing with label switching in mixture models. *J. Roy. Statist. Soc.: Ser. B* (*Statist. Methodology*) 62(4):795–809.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc.: Ser. B* (*Statist. Methodology*) 58(1): 267–288.

Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2004) Sparsity and smoothness via the fused lasso. *J. Roy. Statist. Soc.: Ser. B* (*Statist. Methodology*) 67(1):91–108.

Toubia O, Hauser JR, Simester DI (2004) Polyhedral methods for adaptive choice-based conjoint analysis. *J. Marketing Res.* 41(1):116–131.

Toubia O, Simester DI, Hauser JR, Dahan E (2003) Fast polyhedral adaptive conjoint estimation. *Marketing Sci.* 22(3): 273–303.

Vapnik V (1998) *Statistical Learning Theory* (Wiley, New York).

Vriens M, Wedel M, Wilms T (1996) Metric conjoint segmentation methods: A Monte Carlo comparison. *J. Marketing Res.* 33(1):73–85.

Wahba G (1990) *Spline Models for Observational Data* (SIAM, Philadelphia).

Wittink DR, Cattin P (1989) Commercial use of conjoint analysis: An update. *J. Marketing* 53(3):91–96.

Yuan M, Lin Y (2005) Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc.: Ser. B* (*Statist. Methodology*) 68(1):49–67.