Management Science

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

On Theoretical and Empirical Aspects of Marginal Distribution Choice Models

Vinit Kumar Mishra, Karthik Natarajan, Dhanesh Padmanabhan, Chung-Piaw Teo, Xiaobo Li

MANAGEMENT SCIENCE
Vol. 60, No. 6, June 2014, pp. 1511–1531
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2014.1906

On Theoretical and Empirical Aspects of Marginal Distribution Choice Models

Vinit Kumar Mishra
Department of Business Analytics, University of Sydney Business School, New South Wales 2006, Australia,
vinit.mishra@sydney.edu.au

Karthik Natarajan
Engineering Systems and Design, Singapore University of Technology and Design, Singapore 138682,
natarajan_karthik@sutd.edu.sg

Dhanesh Padmanabhan
General Motors Research and Development–India Science Lab, Bangalore 560066, India,
dhanesh.padmanabhan@gm.com

Chung-Piaw Teo
Department of Decision Sciences, National University of Singapore Business School, Singapore 117591,
bizteocp@nus.edu.sg

Xiaobo Li
Industrial and Systems Engineering, University of Minnesota, Minneapolis, Minnesota 55455,
lixx3195@umn.edu

In this paper, we study the properties of a recently proposed class of semiparametric discrete choice models (referred to as the marginal distribution model (MDM)), by optimizing over a family of joint error distributions with prescribed marginal distributions. Surprisingly, the choice probabilities arising from the family of generalized extreme value models of which the multinomial logit model is a special case can be obtained from this approach, despite the difference in assumptions on the underlying probability distributions. We use this connection to develop flexible and general choice models to incorporate consumer and product level heterogeneity in both partworths and scale parameters in the choice model. Furthermore, the extremal distributions obtained from the MDM can be used to approximate the Fisher's information matrix to obtain reliable standard error estimates of the partworth parameters, without having to bootstrap the method. We use simulated and empirical data sets to test the performance of this approach. We evaluate the performance against the classical multinomial logit, mixed logit, and a machine learning approach that accounts for partworth heterogeneity. Our numerical results indicate that MDM provides a practical semiparametric alternative to choice modeling.

Keywords: discrete choice model; convex optimization; machine learning; applied probability
History: Received April 1, 2012; accepted December 18, 2013, by Eric Bradlow, special issue on business analytics. Published online in *Articles in Advance* May 5, 2014.

1. Introduction

Conjoint analysis is used in practice to determine how consumers choose among products and services, based on the utility maximization framework. It allows companies to decompose consumers preferences into partworths (or utilities), associated with each level of each attribute of products and services. Since the early work of McFadden (1974) on the logit based choice model, and the subsequent introduction of the generalized extreme value models (see McFadden 1977, 1978), discrete choice models have been used extensively in many areas in economics, marketing, and transportation research.

In the simplest discrete choice model, the utility of a consumer $i \in \mathcal{I} = \{1, 2, \ldots, I\}$ for an alternative $k \in \mathcal{K} = \{1, 2, \ldots, K\}$ is decomposed into a sum of a deterministic and a random utility given by $\tilde{U}_{ik} = V_{ik} + \tilde{\epsilon}_{ik}$. The first component V_{ik} is a function of the consumers preference weights $\boldsymbol{\beta}_i$, expressed as $V_{ik}(\boldsymbol{\beta}_i)$, that represents the deterministic utility obtained from the observed attributes of the alternative in the choice task. Typically, $\boldsymbol{\beta}_i$ is the vector of the consumer's preference weights (or partworths) for the vector of attributes of alternative k offered to consumer i, denoted by $\mathbf{x}_{ik}$ with $V_{ik}(\boldsymbol{\beta}_i) = \boldsymbol{\beta}_i' \mathbf{x}_{ik}$. The second component $\tilde{\epsilon}_{ik}$ denotes the random and unobservable or idiosyncratic term in the utility.

Prediction of consumer i's choice for alternative k is done by evaluating the choice probability $P_{ik} := P(\tilde{U}_{ik} \geq \max_{l \in \mathcal{K}} \tilde{U}_{il}) = P(V_{ik} + \tilde{\epsilon}_{ik} \geq \max_{l \in \mathcal{K}} (V_{il} + \tilde{\epsilon}_{il}))$.

The computation involves an evaluation of a multidimensional integral, and closed-form solutions are available only for certain classes of distributions. This includes the generalized extreme value (GEV) family, which is derived under the assumption that the error terms follow a generalized extreme value distribution. The multinomial logit (MNL) and nested logit (NestL) models are well-known members of this family. The choice probabilities do not have closed-form solutions for most other distributions.

In practice, an adequate modeling of consumer heterogeneity is important for accurate choice prediction. To account for taste variation among the consumers, the mixed logit (MixL) model (see, e.g., Train 2009, Allenby and Rossi 1999) assumes that the partworth parameter vector $\boldsymbol{\beta}_i$ is sampled from a distribution $\boldsymbol{\beta}_0 + \tilde{\boldsymbol{\epsilon}}_i^a$. In this way, the utility function $\tilde{U}_{ik} = (\boldsymbol{\beta}_0 + \tilde{\boldsymbol{\epsilon}}_i^a)'\mathbf{x}_{ik} + \tilde{\epsilon}_{ik}$ captures consumer taste variation across the attributes. By integrating over the density, the choice probabilities under the MixL model is derived as $P_{ik} = \int P_{ik}(\boldsymbol{\epsilon}_i^a) g(\boldsymbol{\epsilon}_i^a)\, d\boldsymbol{\epsilon}_i^a$, where $P_{ik}(\boldsymbol{\epsilon}_i^a)$ is the choice probability for given $\boldsymbol{\epsilon}_i^a$, and $g(\cdot)$ denote the probability density of $\tilde{\boldsymbol{\epsilon}}_i^a$. For instance, when the error terms $\tilde{\epsilon}_{ik}$ are independent and identically distributed (i.i.d.) Gumbel, the MNL formula applies with

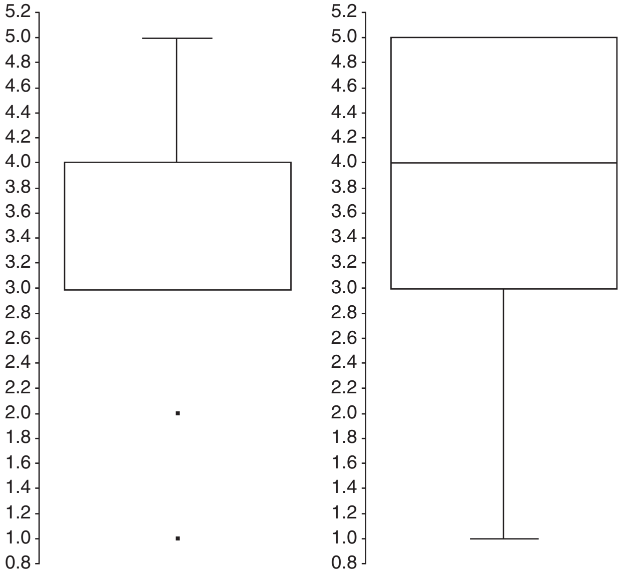
$$P_{ik}(\boldsymbol{\epsilon}_i^a) = \frac{e^{(\boldsymbol{\beta}_0+\boldsymbol{\epsilon}_i^a)'\mathbf{x}_{ik}}}{\sum_{l\in\mathcal{K}} e^{(\boldsymbol{\beta}_0+\boldsymbol{\epsilon}_i^a)'\mathbf{x}_{il}}}.$$

At the same time, $g(\cdot)$ is typically assumed to be a continuous unimodal distribution such as the multivariate normal distribution. The choice probabilities and parameters for the MixL model are then estimated using simulation techniques.

An alternate way to deal with consumer taste variation is to use a regularization technique commonly employed in machine learning. Evgeniou et al. (2007) propose a convex optimization formulation in the discrete choice setting as an alternative to the MixL model. Their approach is based on augmenting the conditional log-likelihood objective with a regularization term that pools information among the consumers and shrinks the individual partworths toward the population mean. This is analogous to estimating the model using the likelihood function $P_{ik}(\boldsymbol{\epsilon}_i^a) g(\boldsymbol{\epsilon}_i^a)$.

In all these approaches, the error terms $\tilde{\epsilon}_{ik}$ are essentially i.i.d. Gumbel, so that the choice probabilities $P_{ik}(\boldsymbol{\epsilon}_i^a)$, can be expressed in closed form. However, this implicitly assumes that the independence of irrelevant alternatives (IIA) property holds at the individual level, even though it does not hold at the population level. Steenburgh and Ainslie (2010) recently pointed out that the MixL model has undesirable properties, in that it restricts the type of substitution patterns that one can expect in the population, as long as the individual IIA property holds. Louviere and Meyer (2007) (see also Fiebig et al. 2010) have also argued that much of the taste heterogeneity encountered in many practical choice tasks can be better described as scale heterogeneity, by assuming that consumers have different scales for the idiosyncratic error terms. They modeled the random $\boldsymbol{\beta}_i$ by $\tilde{\epsilon}_i^a \boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0$ is independent of i, and $\tilde{\epsilon}_i^a$ is a random scale term for each consumer. Note that in these models, the idiosyncratic errors $\tilde{\epsilon}_{ik}$ for the alternatives, after the scaling by ϵ_i^a, are still i.i.d. Gumbel, to ensure tractability of the choice model.

Figure 1 Plots of the Ratings for Two Movies



These assumptions, however, are not realistic for many choice tasks, especially for the case when one of the alternatives is presented as a no-choice or opt-out option. In these cases, the assumption that the idiosyncratic error distribution for the opt-out option is identical to the rest of the product options is often too strong. This assumption is also not valid for many product evaluations. For instance, using the data from GroupLens Research page,[1] the ratings for two randomly selected movies (by 452 and 136 users, respectively) in the data set are as shown in Figure 1. Clearly, the scale of the ratings for the two movies are different, as the spreads of the ratings for the two movies differ. If ratings are indicative of the utilities associated with the movies, it is then essential to account for the difference in scales in the idiosyncratic error for choice modeling in movie selections.

There have been other attempts to investigate the benefits of incorporating heteroskedastic error distributions into choice models. Most of these studies build on the heteroskedastic extreme value (HEV) model introduced by Bhat (1995), based on the assumption of

[1] See http://grouplens.org/datasets/movielens/#attachments (accessed November 2013).

independent extreme value error distributions with nonidentical scale. Since a closed-form expression for the choice probabilities under HEV is not available, Bhat (1995) developed a Gaussian quadrature technique to estimate the parameters of HEV using the maximum likelihood method.

In this paper, we build on a recently proposed class of semiparametric choice models that allow the idiosyncratic error terms to have different scales across products/alternatives and is computationally tractable. In a way, this approach extends the nested logit model, where the scale associated with nests may be different. More importantly, our approach can be combined with the machine learning approach discussed earlier to handle both partworth and scale heterogeneity across consumers and products/attributes. In fact, incorporating regularization terms often speeds up the convergence rate of the estimation procedure. Our approach has the added computational advantage that the calibration problem can be solved using standard nonlinear optimization packages. This approach is along the line of work in the semiparametric and nonparametric literature (see Manski 1975, Farias et al. 2013), in that we do not assume that the distribution of the idiosyncratic error terms are completely specified. For these models, the standard errors of the estimates are often obtained via a bootstrapping approach. In our approach, the marginals in MDM can be used to approximate the Fisher's information matrix to obtain reliable standard error estimates of the partworth parameters directly.

The structure of the rest of the paper and our main contributions are as follows:

In §2, we introduce the marginal distribution model (MDM) and establish its connection with the generalized extreme value model. We show that the GEV choice probabilities can be obtained as an instance of MDM using generalized exponential distributions. In a special case, this implies that MDM with identical exponential distributions generates the MNL choice probabilities. This is surprising because MDM is obtained without the assumption of independence, whereas the MNL model imposes the assumption of independence among the error terms. In fact, we show that under appropriate choice of marginals, there is a one-to-one correspondence between all choice probabilities in the unit simplex and the deterministic components of the utilities. As an illustration, the choice probabilities for a distributionally robust choice model (with known first and second moments for the marginals) can be derived from MDM using t-distributions.

In §3, we study the parameter estimation problem under the MDM using the maximum log-likelihood approach. This estimation problem is known to be convex for only a few special cases. We show that under a linear utility specification, the estimation problem for MDM is convex in partworths under appropriate conditions for several classes of marginal distributions. This includes the MNL and NestL results as special cases. We show further that the method can be used to find asymptotic variance of maximum likelihood estimators under MDM.

In §§4 and 5, we test the performance of MDM using simulated data and empirical conjoint choice data on safety features in automobiles. For companies such as General Motors, understanding consumer preferences for these features is important as it provides key insights on packaging features for current and future vehicles. We use three types of models—MNL with and without heterogeneous partworth preferences, mixed logit with random partworths, and instances of MDM—to analyze the data. We find that despite the nonconvexity of the calibration problem for MDM when both scale and partworth parameters need to be estimated, the optimization problem can be solved fairly efficiently using standard nonlinear optimization packages. The ability to account for scale heterogeneity among consumers using MDM is shown to be useful, because it improves the out-of-sample hit rate by around 20%, in comparison to the simplest MNL model. The out-of-sample predictions in this data set also indicate that MDM with consumer scale heterogeneity performs as well, if not better, than MNL with consumer partworth heterogeneity, despite using much fewer variables and at a fraction of the computational time. We also propose a way to derive standard error estimates for partworths based on the MDM approach, and compare its performance with the standard mixed logit.

2. Choice Prediction Under MDM

In this section, we introduce the marginal distribution model and establish useful properties for the choice probabilities obtained from this model.

2.1. Marginal Distribution Choice Models

Natarajan et al. (2009) recently proposed a semiparametric approach to choice modeling using limited information on the joint distribution of the random utilities. Under this model, the choice prediction is performed as follows. The modeler assumes that the consumer i is utility maximizing while making a choice among the alternatives $k \in \mathcal{K}$ (with utilities U_{ik}) and solves

$$Z(\tilde{\mathbf{U}}_i) = \max\left\{\sum_{k\in\mathcal{K}} \tilde{U}_{ik} y_{ik}\colon\ \sum_{k\in\mathcal{K}} y_{ik} = 1,\quad y_{ik} \in \{0,1\}\ \forall\, k \in \mathcal{K}\right\}. \quad (1)$$

The key difference in the marginal distribution model lies in the evaluation of choice probabilities. Rather than assuming a probability distribution θ of the random utility vector $\tilde{\mathbf{U}}_i = (\tilde{U}_{i1}, \ldots, \tilde{U}_{iK})$, the choice probabilities are evaluated for a probability distribution θ^* of the random utility vector $\tilde{\mathbf{U}}_i$ that satisfies certain prespecified conditions (e.g., fixed marginal distributional information or marginal moment information) and is extremal with respect to the consumer expected welfare. An extremal distribution here refers to the assumption that θ^* is chosen from a set of probability distributions Θ so that it maximizes the consumer expected welfare:

$$\max_{\theta \in \Theta} E_\theta(Z(\tilde{\mathbf{U}}_i)). \tag{2}$$

The extremal distribution is thus given by

$$\theta^* = \arg\max_{\theta \in \Theta} E_\theta(Z(\tilde{\mathbf{U}}_i)). \tag{3}$$

Given marginal distributions in the description of Θ, the framework of "copula modeling" from probability theory (see Sklar 1959, Nelsen 2006) provides a natural way in which univariate marginal distributions can be combined to create multivariate distributions. Besides the popular use of copula in financial risk management, Danaher and Smith (2011) recently identified the flexibility of a copula modeling approach in marketing applications. Examples of copula used in marketing applications include the Gaussian copula (see Danaher and Smith 2011, Seetharman et al. 2005) and the Sarmanov copula (see Schweidel et al. 2008, Park and Fader 2004, Danaher 2007). Our choice of the extremal distribution in (3) gives rise to joint random variables, which is commonly referred to in copula theory literature as countermonotonic random variables (see Nelsen 2006, Weiss 1986).[2]

When Θ denotes the family of probability distributions with prescribed marginals, we obtain the marginal distribution model. When Θ denotes the family of all probability distributions with prescribed first and second marginal moments, we obtain the marginal moment model (MMM). Under the extremal joint distribution θ^* of the random utility vector,

$$\begin{aligned} E_{\theta^*}(Z(\tilde{\mathbf{U}}_i)) &= E_{\theta^*}\left(\sum_{k \in \mathcal{K}} \tilde{U}_{ik} y^*_{ik}(\tilde{\mathbf{U}}_i)\right) \\ &= \sum_{k \in \mathcal{K}} E_{\theta^*}(\tilde{U}_{ik} \mid y^*_{ik}(\tilde{\mathbf{U}}_i) = 1) P_{\theta^*}(y^*_{ik}(\tilde{\mathbf{U}}_i) = 1), \end{aligned}$$

where $y^*_{ik}(\tilde{\mathbf{U}}_i)$ is the optimal value of decision variable y_{ik} in (1), which is random because of random coefficients $\tilde{U}_{ik}$, and $P_{\theta^*}(y^*_{ik}(\tilde{\mathbf{U}}_i) = 1)$ is the choice probability of kth alternative for consumer i under the extremal distribution. For a detailed discussion on these models, the reader is referred to Natarajan et al. (2009), where the model is derived and its application to discrete choice modeling is provided.

For the connection to the persistency problem, see Bertsimas et al. (2006). The key result of Natarajan et al. (2009) for MDM is given in the following theorem.

Theorem 1 (Natarajan et al. 2009). *For consumer i, assume that the marginal distribution $F_{ik}(\cdot)$ of the error term $\tilde{\epsilon}_{ik}$ is a continuous distribution for all $k \in \mathcal{K}$. The following concave maximization problem solves* (2):

$$\max_{\mathbf{P}_i} \left\{ \sum_{k \in \mathcal{K}} \left(V_{ik} P_{ik} + \int_{1-P_{ik}}^{1} F_{ik}^{-1}(t)\, dt \right) : \sum_{k \in \mathcal{K}} P_{ik} = 1, \; P_{ik} \geq 0 \;\forall\, k \in \mathcal{K} \right\} \tag{4}$$

and the choice probabilities under an extremal distribution θ^* of (3) *is the optimal solution vector $\mathbf{P}^*_i$ to* (4).

Under MDM, the optimality conditions yield the choice probabilities as

$$P^*_{ik} = 1 - F_{ik}(\lambda_i - V_{ik}), \tag{5}$$

where the Lagrange multiplier λ_i satisfies the following normalization condition:

$$\sum_{k \in \mathcal{K}} P^*_{ik} = \sum_{k \in \mathcal{K}} (1 - F_{ik}(\lambda_i - V_{ik})) = 1. \tag{6}$$

Natarajan et al. (2009) provide an explicit characterization of the extremal distribution θ^*.[3]

This result has several immediate implications. Take for instance $F_{ik}(\epsilon) = 1 - e^{-\epsilon}$ for $\epsilon \geq 0$. Solving the optimality condition for MDM for (5) and (6), we obtain the MNL choice probabilities:

$$P_{ik} = \frac{e^{V_{ik}}}{\sum_{l \in \mathcal{K}} e^{V_{il}}}.$$

In fact, the formulation in Theorem 1 in this case is exactly the entropy-type utility maximization problem proposed by Anderson et al. (1988) to derive the MNL choice probabilities via a "representative consumer" model. The extremal distribution θ^* obtained from the MDM provides an alternative explanation to the representative consumer model of Anderson et al. (1988).

[2] For $K = 2$, countermonotonic random variables are generated from a proper copula, and for $K > 2$ the countermonotonic random variables are generated from a proper quasi-copula. A detailed discussion on this topic can be found in Nelsen (2006).

[3] This distribution is a finite mixture distribution of the form $\theta^*(\boldsymbol{\epsilon}_i) = \sum_{k \in \mathcal{K}} P^*_{ik} \hat{f}_{ik}(\boldsymbol{\epsilon}_{\mathrm{i}})$, where P^*_{ik} satisfies the conditions (5) and (6) and the density function for the error terms are generated independently from the density functions of the error terms $f_{ik}(\cdot)$ as follows:

$$\hat{f}_{ik}(\boldsymbol{\epsilon}_{\mathrm{i}}) = \frac{f_{ik}(\epsilon_{ik}) I_{\{\epsilon_{ik} \geq \lambda_i - V_{ik}\}}}{1 - F_{ik}(\lambda_i - V_{ik})} \prod_{l \in \mathcal{K}: l \neq k} \frac{f_{il}(\epsilon_{il}) I_{\{\epsilon_{il} \leq \lambda_i - V_{il}\}}}{F_{il}(\lambda_i - V_{il})} \quad \text{for } k \in \mathcal{K}.$$

It is clear from this construction that the joint distribution under MDM is highly correlated.

An important consideration that a modeler faces in the selection of a discrete choice model is that the utility model selected must be capable of generating all possible choice probabilities in the unit simplex. Hofbauer and Sandholm (2002) and Norets and Takahashi (2013) have shown that under mild assumptions on the joint distribution of the error terms, the mapping from the deterministic components of the utilities to the set of choice probabilities is subjective. Hence, any vector of choice probabilities in the unit simplex can be obtained by selecting suitable V_{ik}. We show in the next theorem that under mild assumptions on the marginal density function, MDM also satisfies desirable properties.

Theorem 2. *Set* $V_{i1}=0$. *Assume MDM with error terms* $\tilde{\epsilon}_{ik}$, $k\in\mathcal{K}$ *that have a strictly increasing continuous marginal distribution* $F_{ik}(\cdot)$ *defined either on a semi-infinite support* $[\underline{\epsilon}_{ik},\infty)$ *or an infinite support* $(-\infty,\infty)$. *Let* Δ_{K-1} *be the* $K-1$ *dimensional simplex of choice probabilities*:

$$\Delta_{K-1}=\left\{\mathbf{P}_i=(P_{i1},\dots,P_{iK}):\ \sum_{k\in\mathcal{K}}P_{ik}=1,\ P_{ik}\ge 0\ \forall\, k\in\mathcal{K}\right\}.$$

Let $\Phi(V_{i2},\dots,V_{iK})$: $\Re_{K-1}\to\Delta_{K-1}$ *be a mapping from the deterministic components of the utilities to the choice probabilities under MDM. Then* ϕ *is a bijection between* $\Re_{K-1}$ *and the interior of the simplex* Δ_{K-1}.

Proof. See the appendix.

This result shows that with the error distribution $F_{ik}(\cdot)$ fixed, the deterministic utility component $\mathbf{V}_i$, with V_{i1} fixed at zero, is identifiable from the choice probabilities $\mathbf{P}_i$. In the appendix, we provide additional conditions under which the deterministic utility component and the parameters of the error distribution $F_{ik}(\cdot)$ are simultaneously identifiable from the choice probabilities $\mathbf{P}_i$.

2.2. Distributionally Robust Model with Known First and Second Moments

In a similar spirit, when error terms $\tilde{\epsilon}_{ik}$ have mean zero and variance σ_{ik}^2, but the exact distribution is not known, the choice probabilities under the marginal moment model can be found by solving the following concave maximization problem:

$$\max_{\mathbf{P}_i}\left\{\sum_{k\in\mathcal{K}}\big(V_{ik}P_{ik}+\sigma_{ik}\sqrt{P_{ik}(1-P_{ik})}\big):\ \sum_{k\in\mathcal{K}}P_{ik}=1,\right.$$
$$\left.P_{ik}\ge 0\ \forall\, k\in\mathcal{K}\right\}.\quad(7)$$

In this case, the optimality conditions generate the choice probabilities

$$P_{ik}^*=\frac{1}{2}\left(1+\frac{V_{ik}-\lambda_i}{\sqrt{(V_{ik}-\lambda_i)^2+\sigma_{ik}^2}}\right),\quad(8)$$

where λ_i satisfies the following normalization condition:

$$\sum_{k\in\mathcal{K}}P_{ik}^*=\sum_{k\in\mathcal{K}}\frac{1}{2}\left(1+\frac{V_{ik}-\lambda_i}{\sqrt{(V_{ik}-\lambda_i)^2+\sigma_{ik}^2}}\right)=1.\quad(9)$$

We show that this choice formula can be obtained from MDM using t-distributions as the marginal distributions. Let

$$F_{ik}(\epsilon)=\frac{1}{2}\left(1+\frac{\epsilon}{\sqrt{\sigma_{ik}^2+\epsilon^2}}\right)\quad(10)$$

denote the distribution function of $\tilde{\epsilon}_{ik}$. From the MDM choice probability (5), we have

$$\begin{aligned}P_{ik}&=1-F_{ik}(\lambda_i-V_{ik})\\&=1-\frac{1}{2}\left(1+\frac{\lambda_i-V_{ik}}{\sqrt{\sigma_{ik}^2+(\lambda_i-V_{ik})^2}}\right)\\&=\frac{1}{2}\left(1+\frac{V_{ik}-\lambda_i}{\sqrt{\sigma_{ik}^2+(\lambda_i-V_{ik})^2}}\right).\end{aligned}$$

This is precisely the choice probability obtained from solving the MMM.

2.3. Generalized Extreme Value Model

The MDM choice probabilities as given by (5) and (6) are quite general in the sense that different choices of the marginal distributions F_{ik} of error terms $\tilde{\epsilon}_{ik}$ leads to different choice probabilities. We show next that this model can be related to the GEV model, namely, all GEV probabilities can be obtained from MDM.

Suppose

$$F_{ik}(\epsilon)=1-e^{-\epsilon}G_{ik}(e^{V_{i1}},\dots,e^{V_{iK}},\boldsymbol{\gamma})$$
$$\text{for }\epsilon\ge\ln(G_{ik}(e^{V_{il}},\dots,e^{V_{iK}},\boldsymbol{\gamma})).\quad(11)$$

Note that $F_{ik}(\cdot)$ is a valid distribution function for $\tilde{\epsilon}_{ik}$ under the assumptions on $G_{ik}(\cdot)$ listed in McFadden (1978), and $\boldsymbol{\gamma}$ are given parameters for the distributions. In this case, the Lagrange multiplier satisfies the following condition:

$$\begin{aligned}\sum_{k\in\mathcal{K}}P_{ik}&=\sum_{k\in\mathcal{K}}(1-F_{ik}(\lambda_i-V_{ik}))\\&=e^{-\lambda_i}\times\sum_{k\in\mathcal{K}}e^{V_{ik}}G_{ik}(e^{V_{i1}},\dots,e^{V_{iK}},\boldsymbol{\gamma})=1.\end{aligned}$$

Solving this equation, we get

$$\lambda_i=\ln\left(\sum_{k\in\mathcal{K}}e^{V_{ik}}G_{ik}(e^{V_{i1}},\dots,e^{V_{iK}},\boldsymbol{\gamma})\right).$$

The consumer i's choice probability for alternative k is then given by

$$P_{ik}=\frac{e^{V_{ik}+\ln G_{ik}(e^{V_{i1}},\dots,e^{V_{iK}},\boldsymbol{\gamma})}}{\sum_{l\in\mathcal{K}}e^{V_{il}+\ln G_{il}(e^{V_{i1}},\dots,e^{V_{iK}},\boldsymbol{\gamma})}}.\quad(12)$$

This is exactly the choice probabilities obtained from the GEV model. Note that

$$\frac{P_{ik}}{P_{il}}=\frac{e^{V_{ik}}}{e^{V_{il}}}\ \frac{G_{ik}(e^{V_{i1}},\dots,e^{V_{iK}},\boldsymbol{\gamma})}{G_{il}(e^{V_{i1}},\dots,e^{V_{iK}},\boldsymbol{\gamma})}.$$

Hence, by specifying appropriate choices of $G_{ik}(\cdot)$, we can overcome the limitation of the IIA property inherent in MNL. For instance, when $G_{ik}(e^{V_{i1}}, \ldots, e^{V_{iK}}, \boldsymbol{\gamma}) = \sum_{m\in\mathcal{K}} \gamma_{km} V_{im}$ for all k, we have

$$\frac{P_{ik}}{P_{il}} = \frac{e^{V_{ik}}}{e^{V_{il}}} \frac{\sum_m \gamma_{km} V_{im}}{\sum_m \gamma_{lm} V_{im}}.$$

If $\gamma_{km} > \gamma_{lm}$, then whenever V_{im} increases, P_{ik}/P_{il} increases too. This captures the substitution pattern for the case when product m takes away more shares from product l then from product k. Hence, the GEV models developed this way need not satisfy the IIA property.

3. Estimation Under MDM

In this section, we provide results on the parameter estimation problem under the MDM by exploiting first-order optimality conditions. Since the choice probabilities of several classical choice models can be recovered from MDM, this provides a new approach to calibrate the parameters in these models.

3.1. Convex Calibration Under MDM

McFadden (1974) showed that the maximum log-likelihood problem under MNL is a convex optimization problem. For the nested logit model, Daganzo and Kusnic (1993) showed that the problem is convex in partworth parameters for a choice of scale parameters if the mean utility is linear in the partworths. Note that in the NestL model, scale parameters are also estimated and the problem is not jointly convex in partworth parameters and scale parameters.

In the following theorem we present a general convexity result for MDM. We assume that the population is homogeneous (i.e., consumer partworths $\boldsymbol{\beta}_i = \boldsymbol{\beta}$), and $V_{ik}(\boldsymbol{\beta})$ is affine in $\boldsymbol{\beta}$. The calibration problem is to estimate the location parameters $\boldsymbol{\beta}$.

Theorem 3. *The maximum log-likelihood problem under MDM can be formulated as a convex optimization problem, when the tail distribution $\bar{F}_{ik}(x) := 1 - F_{ik}(x)$ of the error terms satisfies the following two conditions: for x in the domain,* (i) *$\bar{F}_{ik}(x)$ is log concave and* (ii) *$\bar{F}_{ik}(x)$ is convex.*

Proof. The maximum log-likelihood problem under the MDM assuming homogeneous consumer partworths $\boldsymbol{\beta}_i = \boldsymbol{\beta}$ can be formulated as the optimization problem

$$\begin{aligned} \max_{\boldsymbol{\lambda}, \boldsymbol{\beta}} \quad & \sum_{i\in\mathcal{I}} \sum_{k\in\mathcal{K}} z_{ik} \ln(1 - F_{ik}(\lambda_i - \boldsymbol{\beta}'\mathbf{x}_{ik})) \\ \text{s.t.} \quad & \sum_{k\in\mathcal{K}} (1 - F_{ik}(\lambda_i - \boldsymbol{\beta}'\mathbf{x}_{ik})) = 1, \quad i \in \mathcal{I}, \end{aligned} \tag{13}$$

where $z_{ik} = 1$ if the consumer $i \in \mathcal{I}$ chooses alternative $k \in \mathcal{K}$ and 0 otherwise. It is easy to see that the constraints in (13) can be relaxed to

$$\sum_{k\in\mathcal{K}} (1 - F_{ik}(\lambda_i - \boldsymbol{\beta}'\mathbf{x}_{ik})) \le 1, \tag{14}$$

since both the objective function and the left-hand side of the unique constraint involving λ_i are decreasing in λ_i. In the optimal solution, λ_i is hence chosen to be the value such that

$$\sum_{k\in\mathcal{K}} (1 - F_{ik}(\lambda_i - \boldsymbol{\beta}'\mathbf{x}_{ik})) = 1.$$

Under the assumption that $\bar{F}_{ik}(x) = 1 - F_{ik}(x)$ is both log concave and convex, the MLE in (13), with the constraints replaced by (14) is a convex optimization problem. □

Note that $\bar{F}_{ik}(x)$ is log concave if $f_{ik}(x)$ is log concave. Furthermore, convexity of $\bar{F}_{ik}(x)$ implies that $f_{ik}(x)$, and hence $\log(f_{ik}(x))$, is nonincreasing. Hence, any density function with $\log(f(x))$ concave and nonincreasing will satisfy the conditions of the theorem. This includes $f(x) = e^{-x}$ as a special case, which corresponds to the well-known MNL choice model.

3.2. Nested Logit

In the classical nested logit model, we have the GEV function $G_{ik}(y_1, \ldots, y_K, \boldsymbol{\gamma} = \{\theta_r\}_r) = \sum_{r=1}^{R} (\sum_{k\in B_r} y_r^{1/\theta_r})^{\theta_r}$, where the variables are partitioned into $r = 1, \ldots, R$ blocks, each with B_r elements. The model is known to be consistent with utility maximizing behavior for $\theta \in (0, 1]$. Let $B_r(k)$ denote the block containing element k. The estimation problem for this case reduces to

$$\begin{aligned} \text{(NestL)} \quad \max_{\boldsymbol{\lambda}, \boldsymbol{\beta}} \quad & \sum_{i\in\mathcal{I}} \sum_{k\in\mathcal{K}} z_{ik} \Bigg(-\lambda_i + \boldsymbol{\beta}'\mathbf{x}_{ik} \\ & + \ln\Bigg(\Bigg(\sum_{l\in B_r(k)} e^{(1/\theta_r)(\boldsymbol{\beta}'\mathbf{x}_{il} - \boldsymbol{\beta}'\mathbf{x}_{ik})} \Bigg)^{\theta_r - 1} \Bigg) \Bigg) \\ \text{s.t.} \quad & \sum_{k\in\mathcal{K}} \Bigg(\sum_{l\in B_r(k)} e^{(1/\theta_r)(\boldsymbol{\beta}'\mathbf{x}_{il} - \boldsymbol{\beta}'\mathbf{x}_{ik})} \Bigg)^{\theta_r - 1} \\ & \times e^{-\lambda_i + \boldsymbol{\beta}'\mathbf{x}_{ik}} \le 1, \quad i \in \mathcal{I}. \end{aligned} \tag{15}$$

Using the theory of constrained optimization under MDM, this approach provides an alternative proof of the result in Daganzo and Kusnic (1993).

Theorem 4. *The estimation problem* (15) *for nested logit is a convex optimization problem if $\theta_r \in (0, 1]$.*

Proof. See the appendix.

3.3. Marginal Exponential Model

As a generalization, we consider a heterogeneous version of MDM in this section where the exponential marginal distributions potentially have different scale parameters. This model is inspired by the heteroskedastic extreme value model of Bhat (1995). Consider an instance of MDM, which we call the marginal exponential model (MEM) where the scale parameters for each alternative for each consumer is given by $\alpha_{ik} > 0$ (possibly different) and the marginal distribution is given as

$$F_{ik}(\epsilon) = 1 - e^{-\alpha_{ik}\epsilon} \quad \text{for } \epsilon \ge 0. \tag{16}$$

Thus, MEM captures heteroskedasticity in the error terms as in the HEV model. For given scale parameters α_{ik}, the estimation problem under MEM reduces to the following convex formulation:

$$\text{(MEM)} \quad \max_{\boldsymbol{\lambda},\boldsymbol{\beta}} \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}} z_{ik}\alpha_{ik}(\boldsymbol{\beta}'\mathbf{x}_{ik}-\lambda_i)$$
$$\text{s.t.} \quad \sum_{k\in\mathcal{K}} e^{\alpha_{ik}(\boldsymbol{\beta}'\mathbf{x}_{ik}-\lambda_i)} \le 1, \quad i\in\mathcal{I}. \tag{17}$$

However, when the scale parameters also need to be estimated, the maximum log-likelihood problem is no longer jointly convex in the $(\boldsymbol{\beta},\boldsymbol{\lambda})$ and $\boldsymbol{\alpha}$ variables. In this case, the problem is formulated as

$$\max_{\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\alpha}} \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}} z_{ik}\alpha_{ik}(\boldsymbol{\beta}'\mathbf{x}_{ik}-\lambda_i)$$
$$\text{s.t.} \quad \sum_{k\in\mathcal{K}} e^{\alpha_{ik}(\boldsymbol{\beta}'\mathbf{x}_{ik}-\lambda_i)} = 1, \quad i\in\mathcal{I} \tag{18}$$
$$\boldsymbol{\alpha}\in\mathscr{C}.$$

For model identification and to avoid overfitting, it is necessary to restrict the possible values of the scale parameters $\boldsymbol{\alpha}$ to be chosen from an appropriate set $\mathscr{C}$.

3.4. Handling Partworth and Scale Heterogeneity

A key issue in discrete choice modeling is to handle heterogeneity in the data. Evgeniou et al. (2007) proposed a class of convex optimization formulations in the discrete choice setting to handle this issue, which can be used as an alternative to the mixed logit model (Allenby and Rossi 1999). Their technique is inspired by regularization techniques from statistics. By relaxing the assumption that the consumers come from a homogeneous population with a common $\boldsymbol{\beta}$, Evgeniou et al. (2007) allow for consumers to have their individual partworths $\boldsymbol{\beta}_i$. We refer to the model in Evgeniou et al. (2007) as β-HetMNL. For β-HetMNL, the log-likelihood objective is supplemented with a regularization term that pools information among the consumers and shrinks the individual partworths toward the population mean. The convex optimization problem for β-HetMNL model is formulated as

(β-HetMNL)

$$\max_{\boldsymbol{\beta}_i,\boldsymbol{\beta}_0,\mathbf{D}} \left\{ \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}} z_{ik}\ln\left(\frac{e^{\boldsymbol{\beta}_i'\mathbf{x}_{ik}}}{\sum_{l\in\mathcal{K}} e^{\boldsymbol{\beta}_i'\mathbf{x}_{il}}}\right) - \gamma\sum_{i\in\mathcal{I}}(\boldsymbol{\beta}_i-\boldsymbol{\beta}_0)\mathbf{D}^{-1}(\boldsymbol{\beta}_i-\boldsymbol{\beta}_0)' \right\},$$
$$\mathbf{D}\succeq 0,\ \operatorname{trace}(\mathbf{D}) = T, \tag{19}$$

where the parameter $\gamma \ge 0$ captures the trade-off between the log-likelihood criterion and the regularization term. For a fixed value of γ, this problem is convex. The value of γ is set with cross-validation techniques that could possibly use either out-of-sample data or alternatively bootstrapping methods with in-sample data. The matrix variable $\mathbf{D}$ is constrained to be a positive semidefinite matrix with trace fixed to an arbitrary constant T (say 1). This matrix is related to the covariance matrix of partworths. A natural extension is to replace the MNL choice probabilities in the above formulation using choice probabilities from MDM. In particular, if we consider the setting where the scale parameters vary across different alternatives with $\alpha_{ik} = \alpha_k$ for each consumer, we can replace the MNL model with MEM to obtain the following β-HetMEM model:

(β-HetMEM)

$$\max_{\boldsymbol{\lambda},\boldsymbol{\beta}_i,\boldsymbol{\beta}_0,\alpha_k,\mathbf{D}} \left\{ \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}} z_{ik}\alpha_k(\boldsymbol{\beta}_i'\mathbf{x}_{ik}-\lambda_i) - \gamma\sum_{i\in\mathcal{I}}(\boldsymbol{\beta}_i-\boldsymbol{\beta}_0)\mathbf{D}^{-1}(\boldsymbol{\beta}_i-\boldsymbol{\beta}_0)' \right\} \tag{20}$$
$$\text{s.t.} \quad \sum_{k\in\mathcal{K}} e^{\alpha_k(\boldsymbol{\beta}_i'\mathbf{x}_{ik}-\lambda_i)} = 1, \quad i\in\mathcal{I},$$
$$\mathbf{D}\succeq 0,\ \operatorname{trace}(\mathbf{D}) = T,$$
$$\boldsymbol{\alpha}\in\mathscr{C}.$$

The β-HetMEM model has K additional scale parameters in α_k compared with the β-HetMNL model and it overcomes the constant scale assumption across alternatives in β-HetMNL. When the scale parameters are identical for different k, it reduces to the β-HetMNL model.

A key feature of MEM is that it has additional flexibility to capture heterogeneity in the scale parameters at the consumer and attribute level. The importance of incorporating scale heterogeneity in the random component of the utilities has been observed by several researchers including Bhat (1995), Louviere (1988), Louviere and Swait (2010), Louviere et al. (2002), Salisbury and Feinberg (2010), and Fiebig et al. (2010). We propose a new class of choice models with scale heterogeneity based on MEM to address this issue. Our model is inspired by the optimization formulation of Evgeniou et al. (2007), but differs in that we assume that $\boldsymbol{\beta}$ is common across the consumers, but allow for the scale parameters to have different values for consumers across products/alternatives. We term this model the α-HetMEM model. The optimization problem under α-HetMEM is formulated as

(α-HetMEM)

$$\max_{\boldsymbol{\lambda},\boldsymbol{\beta},\alpha_{ik},\alpha_k} \left\{ \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}} z_{ik}\alpha_{ik}(\boldsymbol{\beta}'\mathbf{x}_{ik}-\lambda_i) - \gamma\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}}(\alpha_{ik}-\alpha_k)^2 \right\} \tag{21}$$
$$\text{s.t.} \quad \sum_{k\in\mathcal{K}} e^{\alpha_{ik}(\boldsymbol{\beta}'\mathbf{x}_{ik}-\lambda_i)} = 1, \quad i\in\mathcal{I},$$
$$\boldsymbol{\alpha}\in\mathscr{C}.$$

The parameter γ captures the trade-off between the log-likelihood criterion and the regularization term for the scale parameters. Once again for model identification, we add in constraints that require $\boldsymbol{\alpha}$ to be chosen from a set $\mathscr{C}$.

Note that (α-HetMEM) has significantly less parameters than (β-HetMEM) and (β-HetMNL) if the number of alternatives is much smaller than the number of attributes defining the alternatives, i.e., $|\{\alpha_{ik}\}| \ll |\{\boldsymbol{\beta}_i\}|$. In the data set obtained from General Motors (see §5) this is indeed the case since the choice tasks involve consumers selecting one from a few safety feature packages that are created from combinations of many attributes. Our computational results with the empirical data set from General Motors also indicates that (α-HetMEM) can in this case explain much of the taste variation with far fewer variables, as compared with (β-HetMEM) and (β-HetMNL). This helps confirm the usefulness of capturing scale heterogeneity in choice models.

3.5. Asymptotic Variance of the Maximum Log-Likelihood Estimators (MLE)

In this section, we derive a methodology to obtain standard error estimates of partworth parameter in MDM. We assume that the population is homogeneous in consumer partworths. Let $\boldsymbol{\beta}^*$ be the maximum log-likelihood estimator and $\boldsymbol{\beta}_0$ be the true partworth parameter vector. From the theory of maximum likelihood estimation, it is well known that under regularity[4] conditions, the maximum likelihood estimator has the following asymptotic properties:

(a) *Consistency*. $\boldsymbol{\beta}^*$ converges in probability to $\boldsymbol{\beta}_0$

(b) *Asymptotic normality*. As the sample size increases, the distribution of $\boldsymbol{\beta}^*$ approaches a Gaussian random vector with mean $\boldsymbol{\beta}_0$ and covariance matrix $[I(\boldsymbol{\beta}_0)]^{-1}$, where $I(\boldsymbol{\beta}_0) = -E_0[\nabla^2_{\boldsymbol{\beta}_0} \ln L(\boldsymbol{\beta}_0)]$, with $\ln L(\boldsymbol{\beta}_0)$ denoting the log-likelihood value at $\boldsymbol{\beta}_0$. When the form of the information matrix $I(\boldsymbol{\beta}_0)$ is not available, the following estimator of the information matrix can be used:

$$\hat{I}(\boldsymbol{\beta}^*) = -[\nabla^2_{\boldsymbol{\beta}} \ln L(\boldsymbol{\beta})]_{\hat{\boldsymbol{\beta}}^*}. \tag{22}$$

We assume that the log-likelihood function in (13) satisfies the standard regularity conditions. We now evaluate second-order partial derivatives of the log-likelihood function under the MDM at the MLE. For the models such as mixed logit and multinomial probit, the evaluation of these partial derivatives is done numerically by manipulating the first-order derivatives of the simulated log likelihood. The estimation is not only computationally challenging, but is also prone to errors. In the case of MDM, analytical expressions can be derived for the second-order partial derivatives. The maximum likelihood problem (13) involves normalization constraints, and additional Lagrange multipliers that need to be accounted for in deriving analytical expressions for second-order partial derivatives. The key step involves deriving partial derivatives of λ_i with respect to the MLE using the normalization constraints. A step-wise method for finding the analytical expressions of second-order partial derivatives for general distribution functions F_{ik} is outlined next.

We present the method when partworth only parameters $\boldsymbol{\beta}$ are estimated. The log likelihood is

$$\ln L(\boldsymbol{\beta}) = \sum_{i\in\mathscr{I}} \sum_{k\in\mathscr{K}} z_{ik} \ln(1 - F_{ik}(\lambda_i - V_{ik}(\boldsymbol{\beta}^*))), \tag{23}$$

where the variables λ_i are the solutions to the equation

$$\sum_{k\in\mathscr{K}} (1 - F_{ik}(\lambda_i - V_{ik}(\boldsymbol{\beta}))) = 1 \quad \forall\, i \in \mathscr{I}. \tag{24}$$

The following steps can be used to evaluate the second-order partial derivatives of the log-likelihood function at MLE $\boldsymbol{\beta}^*$.

Step 1. For each $i \in \mathscr{I}$ differentiate constraint (24) with respect to $\boldsymbol{\beta}$. We get the following expression:

$$\nabla_{\boldsymbol{\beta}} \lambda_i = \frac{\sum_{k\in\mathscr{K}} f_{ik}(\lambda_i - V_{ik}) \nabla_{\boldsymbol{\beta}} V_{ik}}{\sum_{k\in\mathscr{K}} f_{ik}(\lambda_i - V_{ik})}.$$

Note that V_{ik} is a function of $\boldsymbol{\beta}$ and f_{ik} are density functions of F_{ik}. The above vector can be differentiated again to find analytical expressions for $\nabla^2_{\boldsymbol{\beta}} \lambda_i$.

Step 2. For $i \in \mathscr{I}$ and $k \in \mathscr{K}$, choice probabilities are $P_{ik} = 1 - F_{ik}(\lambda_i - V_{ik})$. Differentiating these probabilities with respect to $\boldsymbol{\beta}$, we get the gradient

$$\nabla_{\boldsymbol{\beta}} P_{ik} = -f_{ik}(\lambda_i - V_{ik})(\nabla_{\boldsymbol{\beta}} \lambda_i - \nabla_{\boldsymbol{\beta}} V_{ik}).$$

Differentiating again, we get the matrix of second partial derivatives as

$$\begin{aligned} \nabla^2_{\boldsymbol{\beta}} P_{ik} = -\big[& f'_{ik}(\lambda_i - V_{ik})(\nabla_{\boldsymbol{\beta}} \lambda_i - \nabla_{\boldsymbol{\beta}} V_{ik})(\nabla_{\boldsymbol{\beta}} \lambda_i - \nabla_{\boldsymbol{\beta}} V_{ik})' \\ & + f_{ik}(\lambda_i - V_{ik})(\nabla^2_{\boldsymbol{\beta}} \lambda_i - \nabla^2_{\boldsymbol{\beta}} V_{ik})\big]. \end{aligned}$$

These expressions can easily be derived using first- and second-order derivatives of λ_i found in Step 1. Note that, in the case when V_{ik} are linear in parameters, $\nabla^2_{\boldsymbol{\beta}} V_{ik} = 0$.

[4] Details of these regularity conditions can be found in Greene (2011) and references provided therein. These conditions would typically require at least the existence of MLE in the interior of feasibility set, and a well-behaved $1 - F_{ik}(\lambda_i - V_{ik}(\boldsymbol{\beta}))$, which is continuous and twice differentiable as a function of estimated parameters etc. Given that MDM replicates GEV of which nested logit and MNL are special cases, for a large class of extreme choice distributions under MDM, consistency and efficiency of MLE follows for these models (see, e.g., McFadden 1974 for MNL and Brownstone and Small 1989 for nested logit).

Step 3. The final step involves differentiating the log-likelihood (23) with respect to $\boldsymbol{\beta}$ twice to find the expression for second-order partial derivatives of log likelihood

$$\nabla^2_{\boldsymbol{\beta}} \ln L(\boldsymbol{\beta}) = \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}} \frac{z_{ik}}{P_{ik}}\left(\nabla^2_{\boldsymbol{\beta}}P_{ik} - \frac{\nabla_{\boldsymbol{\beta}}P_{ik}\nabla_{\boldsymbol{\beta}}P'_{ik}}{P_{ik}}\right).$$

Using expressions derived in Step 2, it is possible to evaluate this at the MLE $\boldsymbol{\beta}^*$.

These values can be thus used to find standard errors of estimators in MDM. In our computational experiments, with additional constraints imposed on the scale parameters, we found that standard error estimates of the scale parameters were much less reliable. Hence, we restrict our attention to finding standard error estimates for the $\boldsymbol{\beta}$ parameters in this paper.

4. Computational Experiments Using Simulated Data

In this section, we test for the empirical identification of the partworth and scale parameters in MEM using a small-scale simulation study with alternative specific scale parameters $\alpha_{ik} = \alpha_k$. Consider a choice set with alternatives indexed as $\mathcal{K} = \{1, 2, 3, 4\}$, where alternative 4 is the outside option. The utility of the four alternatives for consumer i is defined as

$$\tilde{U}_{ik} = \text{ASC}_k + \beta_1 x_{ik1} + \beta_2 x_{ik2} + \beta_3 x_{ik3} + \tilde{\epsilon}_{ik} \quad \forall\, i \in \mathcal{I},\ \forall\, k \in \{1, 2, 3\},$$

$$\tilde{U}_{i4} = \tilde{\epsilon}_{i4} \quad \forall\, i \in \mathcal{I},\ \forall\, k \in \mathcal{K}.$$

The constants ASC_k denote the alternative specific constants for the first three alternatives and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ denotes the partworths for three attributes common to the alternatives. The attribute values $\mathbf{x}_{ik} = (x_{ik1}, x_{ik2}, x_{ik3})$ were generated from random independent draws uniformly chosen in the interval $[0, 1]$. The choice tasks were simulated for 15 instances of randomly generated parameters as follows: Alternative specific constants ASC_k were chosen randomly in the range $[-1, 1]$, partworths $\boldsymbol{\beta}$ were generated randomly in the range $[-2, 2]$, scale parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ were generated randomly in the range $[0.04, 2]$. The scale parameter for the outside option was set to 1. For a given instance of the parameters, we simulated the error terms from the extremal distribution in MEM (see §2) for $I = 5{,}000$, $I = 10{,}000$, $I = 15{,}000$ and $I = 20{,}000$ consumers. Using the choices made by the consumers in the simulation, we estimated the parameters using the following log-likelihood estimation model based on (18):

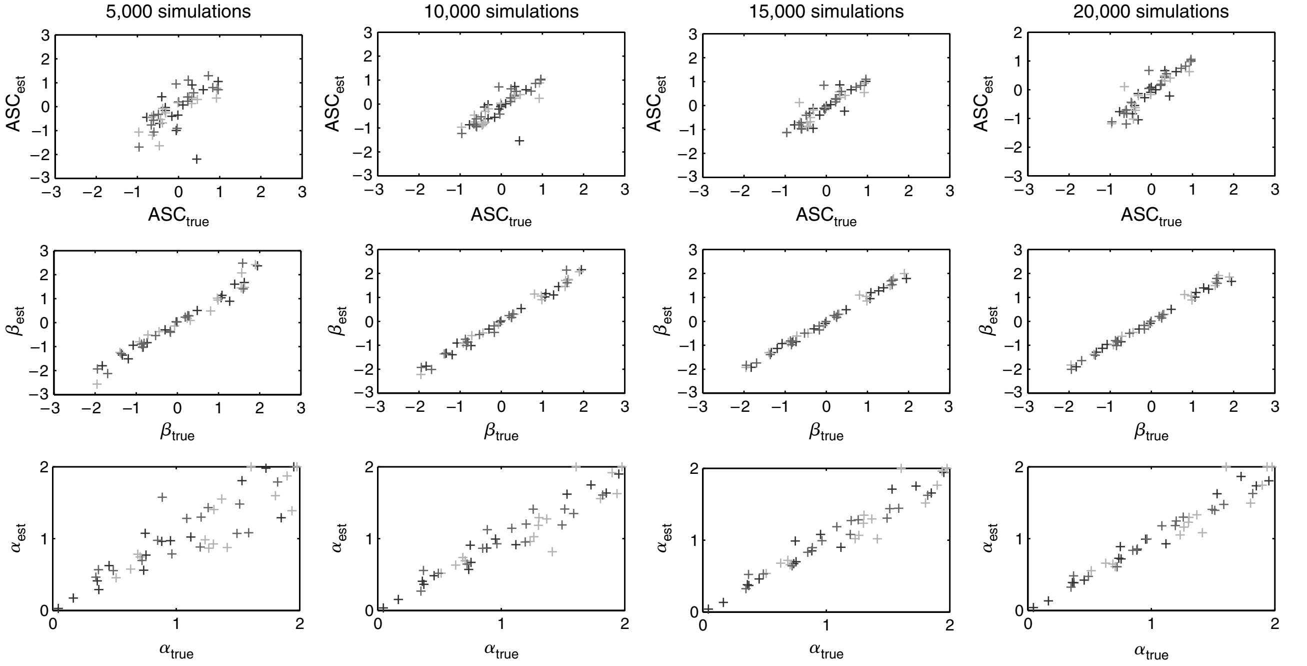
$$\max_{\boldsymbol{\lambda},\,\text{ASC},\,\boldsymbol{\beta},\,\boldsymbol{\alpha}} \left\{\sum_{i\in\mathcal{I}}\sum_{k\in\{1,2,3\}} z_{ik}\alpha_k(\text{ASC}_k + \boldsymbol{\beta}'\mathbf{x}_{ik} - \lambda_i) + \sum_{i\in\mathcal{I}} z_{i4}(-\lambda_i)\right\} \tag{25}$$

$$\text{s.t.} \quad \sum_{k\in\{1,2,3\}} e^{\alpha_k(\text{ASC}_k + \boldsymbol{\beta}'\mathbf{x}_{ik} - \lambda_i)} + e^{-\lambda_i} = 1, \quad i \in \mathcal{I},$$

$$0 \le \alpha_k \le 2,\ k \in \{1, 2, 3\}.$$

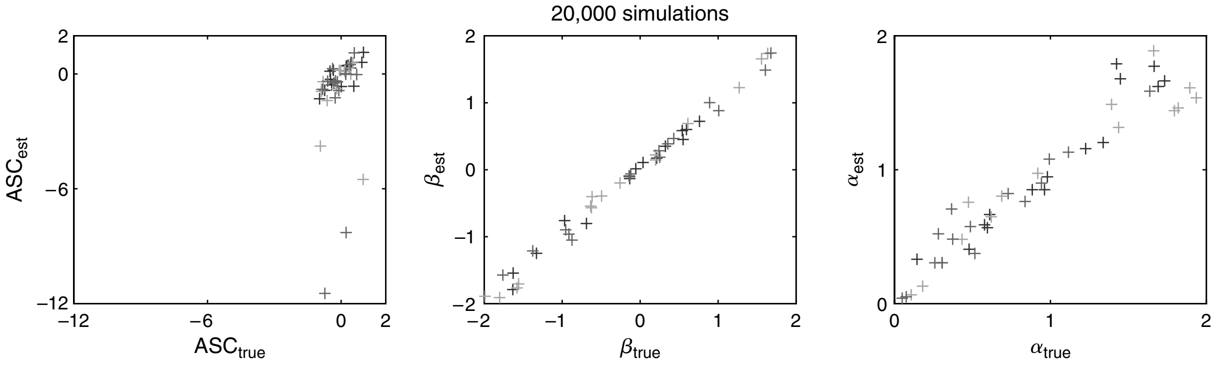
The optimization problems were coded in AMPL and solved using the LOQO solver. The scatter plot for the true and estimated parameters for the 15 sets of parameters are plotted in Figure 2. From Figure 2,

Figure 2 Comparison of True and Estimated Parameters from MEM



5,000 simulations
10,000 simulations
15,000 simulations
20,000 simulations

Figure 3 Comparison of True Parameters in MNL with Different Scale Parameters and Estimated Parameters from MEM



we observe that as the number of simulated choices increases, the estimated parameters clearly get closer to the true parameters. If one knows the true values of the α_k parameters, then from our previous results, the estimation problem is convex, and hence one would be guaranteed to converge to the true solution. Since we also estimate the scale parameters for the alternatives, the optimization problem is nonconvex. The results indicate that while nonconvex, the LOQO solver in conjunction with an appropriate optimization formulation is able to identify the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters fairly well in this example.

To test the robustness of the parameter estimates, we simulated choices by varying the underlying distribution of the error terms. Results in §2 imply that the MEM choice probabilities with exponential marginals is the same as the MNL choice probabilities with Gumbel distributions for the identical scale parameter case. In our simulations, we generate heteroskedastic error distributions using independent Gumbel random variables with $F_{ik}(\epsilon) = e^{-e^{-\alpha_k \epsilon}}$ and simulate choices. Using the optimization formulation for MEM, we then estimate the parameters. The results are provided in Figure 3 for 20,000 simulated choices. The results indicate that the model is fairly effective in identifying the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters despite model misspecification. For the alternative specific constants, we found that in four of the 45 parameters there was a fairly big difference in the actual values of the ASC parameters and the estimated ASC parameters. A more in-depth study of these cases showed that this corresponds to alternatives with very low scale parameters $\alpha_k \leq 0.1$. In this case, the presence of the product term in the estimation model $\alpha_k \text{ASC}_k$ results in more extreme estimates of the alternative specific constant in the optimization formulation.

5. Computational Experiments Using Empirical Data

In this section, we apply the models developed in this paper to empirical data provided by General Motors. The choice-based conjoint data was collected by General Motors to understand consumer's tradeoffs for various vehicle features. We were provided conjoint choice data for a total of 500 consumers. Each consumer performed 19 choice tasks. The data were collected using the CBC/Web system of the choice-based conjoint from Sawtooth Software (2008). There were 19 feature attributes and one "price" attribute. Each attribute has between two to 11 levels. In each choice task, three alternatives (feature packages) were shown based on a partial profile conjoint experiment (see Chrzan and Elrod 1995), where only information on 10 attributes were shown at a time, with the "price" attribute being shown in all tasks. In each choice task, the consumer either chose one of the three alternatives or the outside option. The latter is referred to as the "no choice alternative." (See Figure 4 for a sample choice task.)

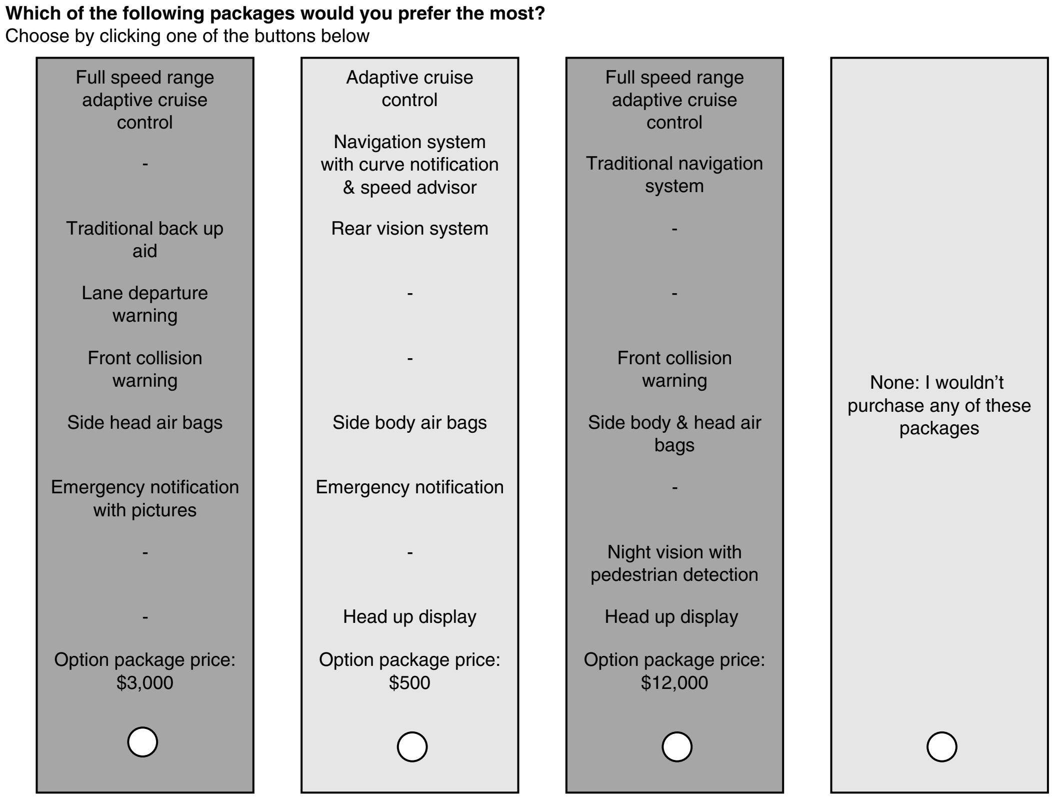
The complete list of product attributes and the attribute level codes are shown in Table 1. Although the full description of each of the attributes and their levels were shown to consumers, we use simple codes to represent them in Table 1. As an illustration, the product attribute NS refers to "navigation system." Level NS1 refers to "standard navigation system," NS2 refers to "navigation system with curve notification," NS3 refers to "navigation system with speed advisor," NS4 refers to "navigation system with curve notification and speed advisor," whereas NS5 refers to "not present." For the "no choice alternative," since there is no information on the attributes, we use zero to denote the mean utility of the alternative.

Apart from the feature package attributes, the data also contained information regarding consumers demographic profile such as age, education, income, gender, and percentage of time the consumer drives at night.

5.1. Classes of Choice Models

We first compare the estimation and prediction results on the General Motors choice data using MNL, NestL,

Figure 4 **A Sample Choice Task**



Which of the following packages would you prefer the most?
Choose by clicking one of the buttons below

Full speed range adaptive cruise control	Adaptive cruise control	Full speed range adaptive cruise control	None: I wouldn't purchase any of these packages
-	Navigation system with curve notification & speed advisor	Traditional navigation system	
Traditional back up aid	Rear vision system	-	
Lane departure warning	-	-	
Front collision warning	-	Front collision warning	
Side head air bags	Side body air bags	Side body & head air bags	
Emergency notification with pictures	Emergency notification	-	
-	-	Night vision with pedestrian detection	
-	Head up display	Head up display	
Option package price: $3,000	Option package price: $500	Option package price: $12,000	
○	○	○	○

and MEM. We compare the three models, both with and without heterogeneity. We estimated the parameters for MNL, NestL, and MEM by maximizing the log-likelihood function. For each respondent $i \in \mathcal{I} = \{1, \ldots, 500\}$, we use the first 12 of the 19 choice tasks performed by the consumer to estimate the parameters of the models. Thus, each consumer i performs $j \in \mathcal{J} = \{1, \ldots, 12\}$ choice tasks in the calibration step. This corresponds to a total of 6,000 in-sample data points (consumer-choice task pairs). The set of available

Table 1 **Attribute and Level Codes**

Serial no.	Attribute name	Attribute code	No. of levels	Level codes
1	Cruise control	CC	3	CC1, CC2, CC3
2	Go notifier	GN	2	GN1,GN2
3	Navigation system	NS	5	NS1, NS2, NS3, NS4, NS5
4	Backup aids	BU	6	BU1, BU2, BU3, BU4, BU5, BU6
5	Front park assist	FA	2	FA1, FA2
6	Lane departure	LD	3	LD1, LD2, LD3
7	Blind zone alert	BZ	3	BZ1, BZ2, BZ3
8	Front collision warning	FC	2	FC1, FC2
9	Front collision protection	FP	4	FP1, FP2, FP3, FP4
10	Rear collision protection	RP	2	RP1, RP2
11	Parallel park aids	PP	3	PP1, PP2, PP3
12	Knee air bags	KA	2	KA1, KA2
13	Side air bags	SC	4	SC1, SC2, SC3, SC4
14	Emergency notification	TS	3	TS1, TS2, TS3
15	Night vision system	NV	3	NV1, NV2, NV3
16	Driver assisted adjustments	MA	4	MA1, MA2, MA3, MA4
17	Low speed braking assist	LB	4	LB1, LB2, LB3, LB4
18	Adaptive front lighting	AF	3	AF1, AF2, AF3
19	Head up display	HU	2	HU1, HU2
20	Price	Price	11	$500, 1,000, 1,500, 2,000, 2,500, 3,000, 4,000, 5,000, 7,500, 10,000, 12,000

alternatives for each choice task is $k \in \mathcal{K} = \{1, \dots, 4\}$, where the fourth alternative is used to represent the "no choice" alternative. The product attribute vector for consumer i in choice task j for alternative k is denoted by $\mathbf{x}_{ijk}$.[5] This is coded as effect-type dummy variables, where each element of $\mathbf{x}_{ijk}$ was set to 1 if the corresponding attribute level was present in the choice task and 0 otherwise. The total number of attributes in this data set was $A = 71$. We use the calibrated model to predict the choices for the remaining seven tasks for each consumer. This corresponded to a total of 3,500 out-of-sample data points (consumer-choice task pairs).

The comparison study was carried out between three classes of models:

(a) *Consumer homogeneity*. In the simplest model, we assume that all the consumers are homogeneous in their partworth preferences and scales of their idiosyncratic errors for different options. Under MNL, the maximum log-likelihood problem is formulated as a convex optimization problem:

$$\text{(MNL)} \quad \max_{\boldsymbol{\beta}} \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} \sum_{k\in\mathcal{K}} z_{ijk} \ln\left(\frac{e^{\boldsymbol{\beta}'\mathbf{x}_{ijk}}}{\sum_{l\in\mathcal{K}} e^{\boldsymbol{\beta}'\mathbf{x}_{ijl}}}\right), \tag{26}$$

where $z_{ijk} = 1$ if the consumer $i \in \mathcal{I}$ in choice task $j \in \mathcal{J}$ chooses the alternative $k \in \mathcal{K}$ and 0 otherwise.

Under MEM, the marginal distribution of each error term is modeled as an exponential distribution: $P(\tilde{\epsilon}_{ijk} \le t) = 1 - e^{-\alpha_k t}$. The parameter α_k in this case corresponds to the scale parameter for each of the four choices $k \in \mathcal{K}$. Since the alternatives shown to the consumers varied with tasks, we consider the simplest possible model where the scale parameters for the first three alternatives are assumed to be identical ($\alpha_1 = \alpha_2 = \alpha_3$), but the scale parameter for the no choice alternative can be different. The maximum log-likelihood problem in this case is

$$\begin{aligned}
\text{(MEM)} \quad \max_{\boldsymbol{\lambda}, \boldsymbol{\beta}, \alpha_k} \quad & \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} \sum_{k\in\mathcal{K}} z_{ijk} \alpha_k (\boldsymbol{\beta}'\mathbf{x}_{ijk} - \lambda_{ij}) \\
\text{s.t.} \quad & \sum_{k\in\mathcal{K}} e^{\alpha_k(\boldsymbol{\beta}'\mathbf{x}_{ijk} - \lambda_{ij})} = 1, \quad i \in \mathcal{I},\ j \in \mathcal{J}, \\
& \sum_{k\in\mathcal{K}} \alpha_k = K, \\
& \alpha_1 = \alpha_2 = \alpha_3, \\
& \alpha_k \ge \text{TOL}, \quad k \in \mathcal{K}.
\end{aligned} \tag{27}$$

The constraint $\sum_{k\in\mathcal{K}} \alpha_k = K$ serves as a normalization constraint. In our experiment, we used $K = 4$ to ensure that the sum of the scale parameters equals the number of alternatives. The constraints $\alpha_k \ge \text{TOL}$ ensure that the scale parameters are strictly positive. In the experiments, we set $\text{TOL} = 0.1$.

In the nested logit model, it is natural to form the first three options as a nest, while the outside option forms another nest. The maximum log-likelihood problem is thus

(NestL)

$$\begin{aligned}
\max_{\boldsymbol{\beta}, \theta} \quad & \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} \left(\sum_{k=1}^{3} z_{ijk} \ln\left(\frac{e^{\boldsymbol{\beta}'\mathbf{x}_{ijk}/\theta} (\sum_{l=1}^{3} e^{\boldsymbol{\beta}'\mathbf{x}_{ijl}/\theta})^{\theta-1}}{(\sum_{l=1}^{3} e^{\boldsymbol{\beta}'\mathbf{x}_{ijl}/\theta})^{\theta} + 1} \right) \right. \\
& \left. - z_{ij4} \ln\left(\left(\sum_{l=1}^{3} e^{\boldsymbol{\beta}'\mathbf{x}_{ijl}/\theta} \right)^{\theta} + 1 \right) \right), \\
\text{s.t.} \quad & \text{TOL} \le \theta \le \text{UB}.
\end{aligned} \tag{28}$$

Note that the upper bound (UB) is imposed since the value of θ must be within a particular range for the model to be consistent with utility-maximizing behavior (see, e.g., Train 2009). In our experiments, we set $UB = 1.8$ and the lower bound TOL to 0.1 as in MEM.

(b) *Consumer partworth heterogenity*. The second class of models uses a regularization approach to model consumer partworth heterogeneity. In β-HetMNL, the log-likelihood objective is supplemented with a regularization term that pools information among the consumers and shrinks the individual partworths toward the population mean. The convex optimization problem for the β-HetMNL model is formulated as

(β-HetMNL)

$$\begin{aligned}
\max_{\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \mathbf{D}} \quad & \left\{ \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} \sum_{k\in\mathcal{K}} z_{ijk} \ln\left(\frac{e^{\boldsymbol{\beta}_i'\mathbf{x}_{ijk}}}{\sum_{l\in\mathcal{K}} e^{\boldsymbol{\beta}_i'\mathbf{x}_{ijl}}} \right) \right. \\
& \left. - \gamma \sum_{i\in\mathcal{I}} (\boldsymbol{\beta}_i - \boldsymbol{\beta}_0) \mathbf{D}^{-1} (\boldsymbol{\beta}_i - \boldsymbol{\beta}_0)' \right\}, \\
\text{s.t.} \quad & \mathbf{D} \succeq 0,\ \text{trace}(\mathbf{D}) = T,
\end{aligned} \tag{29}$$

where the parameter $\gamma \ge 0$ captures the trade-off between the log-likelihood criterion and the regularization term. In our experiments without loss of generality, we set T equal to the number of attributes. For a fixed value of γ, this problem is convex. We found the value of γ using cross-validation with additional out-of-sample data. We vary γ in a large enough interval range to find the value of γ that provides best performance out of sample in terms of log likelihood. The matrix $\mathbf{D}$ is obtained using the sequential optimization approach described in Evgeniou et al. (2007).

An extension of this model is to incorporate partworth heterogeneity into the new MEM.

[5] In the earlier sections, we have assumed that each consumer performs only one choice task, and hence we have omitted the index j in our earlier discussion.

The optimization problem under β-HetMEM is formulated as

$$(\beta\text{-HetMEM})$$

$$\begin{aligned}\max_{\boldsymbol{\lambda},\boldsymbol{\beta}_i,\boldsymbol{\beta}_0,\alpha_k,\mathbf{D}}\ &\Big\{\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} z_{ijk}\alpha_k(\boldsymbol{\beta}_i'\mathbf{x}_{ijk}-\lambda_{ij})\\ &\quad-\gamma\sum_{i\in\mathcal{I}}(\boldsymbol{\beta}_i-\boldsymbol{\beta}_0)\mathbf{D}^{-1}(\boldsymbol{\beta}_i-\boldsymbol{\beta}_0)'\Big\}\\ \text{s.t.}\ &\sum_{k\in\mathcal{K}} e^{\alpha_k(\boldsymbol{\beta}_i'\mathbf{x}_{ijk}-\lambda_{ij})}=1,\quad i\in\mathcal{I},\ j\in\mathcal{J},\\ &\sum_{k\in\mathcal{K}}\alpha_k=4,\\ &\alpha_1=\alpha_2=\alpha_3,\\ &\alpha_k\ge\text{TOL},\quad k\in\mathcal{K},\\ &\mathbf{D}\succeq 0,\ \text{trace}(\mathbf{D})=T.\end{aligned}\tag{30}$$

(c) *Consumer scale heterogenenity*. The third class of models assumes that $\boldsymbol{\beta}$ is common across the consumers but allows for the scale parameter to vary across consumers and alternatives. In our data set, since the first three alternatives vary across different choice tasks, we simplify the model by assuming that $\alpha_{i1}=\alpha_{i2}=\alpha_{i3}$, but allow for the scale of the no choice alternative α_{i4} to be different. We term this model as α-HetMEM. The optimization problem under α-HetMEM is formulated as

$$(\alpha\text{-HetMEM})$$

$$\begin{aligned}\max_{\boldsymbol{\lambda},\boldsymbol{\beta},\alpha_{ik},\alpha_1}\ &\Big\{\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} z_{ijk}\alpha_{ik}(\boldsymbol{\beta}'\mathbf{x}_{ijk}-\lambda_{ij})-\gamma\sum_{i\in\mathcal{I}}(\alpha_{i1}-\alpha_1)^2\Big\}\\ \text{s.t.}\ &\sum_{k\in\mathcal{K}} e^{\alpha_{ik}(\boldsymbol{\beta}'\mathbf{x}_{ijk}-\lambda_{ij})}=1,\quad i\in\mathcal{I},\ j\in\mathcal{J},\\ &\sum_{k\in\mathcal{K}}\alpha_{ik}=4,\quad i\in\mathcal{I},\\ &\alpha_{i1}=\alpha_{i2}=\alpha_{i3},\quad i\in\mathcal{I},\\ &\alpha_{ik}\ge\text{TOL},\quad i\in\mathcal{I},\ k\in\mathcal{K}.\end{aligned}\tag{31}$$

The parameter γ captures the trade-off between the log-likelihood criterion and the regularization term for the scale parameters. Once again, we set γ using cross-validation.

Parallel to the α-HetMEM, one can also allow the scale parameters to vary among different consumers using the classical MNL formula. Interestingly, a straightforward implementation of this scale heterogeneous model does not improve the performance of the model. We can also incorporate scale heterogeneity into the nested logit model. We term this model as θ-HetNest. We implement this model as a benchmark for the α-HetMEM model. The optimization problem under θ-HetNest is formulated as

$$(\theta\text{-HetNest})$$

$$\begin{aligned}\max_{\boldsymbol{\beta},\theta_i,\theta}\ &\Bigg\{\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\Bigg(\sum_{k=1}^{3} z_{ijk}\ln\Bigg(\frac{e^{\boldsymbol{\beta}'\mathbf{x}_{ijk}/\theta_i}\big(\sum_{l=1}^{3}e^{\boldsymbol{\beta}'\mathbf{x}_{ijl}/\theta_i}\big)^{\theta_i-1}}{\big(\sum_{l=1}^{3}e^{\boldsymbol{\beta}'\mathbf{x}_{ijl}/\theta_i}\big)^{\theta_i}+1}\Bigg)\\ &\quad-z_{ij4}\ln\Bigg(\Bigg(\sum_{l=1}^{3}e^{\boldsymbol{\beta}'\mathbf{x}_{ijl}/\theta_i}\Bigg)^{\theta_i}+1\Bigg)\Bigg)-\gamma\sum_{i\in\mathcal{I}}(\theta_i-\theta)^2\Bigg\},\\ \text{s.t.}\ &\text{TOL}\le\theta_i\le\text{UB},\quad i\in\mathcal{I}.\end{aligned}\tag{32}$$

The parameter γ again captures the trade-off between the log-likelihood criterion and the regularization term for the scale parameters.

5.2. Results

We compared the methods using four different criterion:

- *Log likelihood* (LL). The fit log likelihood (in sample) and the predicted log likelihood (out of sample) were used to compare the fit and prediction of the choice models.
- *Akaike information criterion* (AIC). AIC is a measure of the relative goodness of fit of the model and is derived based on information entropy concepts. AIC describes the trade-off between the accuracy and complexity of the model and is defined as $\text{AIC}=2P-\text{LL}$, where P is the number of parameters to estimate in the model, and LL is the log likelihood for the estimated model. A lower value of AIC indicates a more preferred model. Thus, AIC rewards not only the goodness of fit, but also includes a penalty that increases with the number of parameters estimated.
- *Hit rate*. The hit rate is defined as the fraction of choice tasks in which the alternative chosen by the consumer coincides with the alternative with the highest predicted probability. Hit rates were evaluated for both the in-sample and out-of-sample data.
- *Computational time*. The computational time that was required to solve the different optimization models were used to compare the models. The optimization routines were coded in AMPL and solved using the LOQO solver.

The results are provided in Tables 2 and 3. Although we have tested the models for a wide range of γ, for each model, we report only the results with the value of γ selected by cross-validation. Several observations can be made from the results:

(a) The models with homogeneous $\boldsymbol{\beta}$ performs roughly the same in both in-sample and out-of-sample experiments, in terms of LL and hit rate (around 46% to 51%).

(b) Incorporating heterogeneity in the consumer partworths significantly improves the in-sample performance over the homogeneous partworth models—both

Table 2 Fit Results for In-Sample Data

Method	Log likelihood (LL)	Parameters (P)	AIC	Time (seconds)	Hit rate
		Consumer homogeneity			
MNL	−7,218.2283	71	7, 360.2283	7.627	0.4682
MEM	−7,217.6459	72	7,361.6459	44.705	0.4677
NestL	−7,217.8022	72	7,361.8022	40.925	0.4677
		Consumer partworth heterogeneity			
β-HetMNL ($\gamma = 4$)	−3,725.5119	40,612	84,949.5119	211.713	0.8223
β-HetMEM ($\gamma = 4$)	−3,725.6659	40,613	84,951.6659	357.723	0.8241
		Consumer scale heterogeneity MEM			
α-HetMEM ($\gamma = 6$)	−5,457.2608	572	6,601.2608	12.728	0.6035
		Consumer scale heterogeneity nested Logit			
θ-HetNest ($\gamma = 0.7$)	−6,100.1974	572	7,244.1974	54.38	0.5075

Note. Results are indicated for the values of γ selected by out-of-sample cross-validation.

log likelihood and hit rates improves drastically for the β-HetMNL and β-HetMEM models. The in-sample hit rate is as high as 80%, whereas it is 60% in out-of-sample experiments. The performance of β-HetMNL and β-HetMEM are very similar, indicating again that adding consumer invariant scale heterogeneity does not improve the β-HetMNL model. Note that these approaches are more computationally intensive, because of significantly more parameters that need to be estimated for model specification. As suggested in Evgeniou et al. (2007), one also needs to sequentially optimize over $\mathbf{D}$ and $\boldsymbol{\beta}$. Table 4 shows the results of the algorithm in 10 steps of our implementation. We note that a good choice of $\mathbf{D}$ can improve both the in-sample and out-of-sample performances significantly. However, the procedure takes more time to solve nearer the optimal choice of $\mathbf{D}$. The reader is referred to Argyriou et al. (2008) for possible techniques to speed up convergence of these methods. The AIC values are the largest for this approach, because of the large number of parameters needed for the calibration problem.

(c) By incorporating respondent heterogeneity in the scale parameters, we obtain the best performance for out-of-sample experiments. There are much fewer parameters needed in the α-HetMEM model (with 572 scale parameters), compared to β-HetMNL (with more than 40,000 parameters corresponding to customer partworth terms). This is reflected in the AIC criterion where the α-HetMEM outperforms the β-HetMNL and β-HetMEM models significantly. The α-HetMEM also has the best out-of-sample log likelihood with $\text{LL} = -3{,}201.4232$ (for $\gamma = 6$). In terms of hit rates, capturing scale heterogeneity provides similar values to the β-HetMEM and β-HetMNL models in out-of-sample experiments. In the appendix, we consider alternative methods that incorporate scale heterogeneity in the MNL framework. The results indicate that α-HetMEM is still the most preferred model.

(d) In terms of running times, the MNL runs in the least time followed by the α-HetMEM and then the β-HetMNL and β-HetMEM models. This is surprising since α-HetMEM is a nonconvex optimization problem and β-HetMNL is convex. However, this can be partly explained by the fact that α-HetMEM has far fewer variables compared with β-HetMNL. Note that α-HetMEM runs faster than MEM, indicating that the convex regularization terms are effective in speeding up the convergence of the algorithm.

A natural extension based on these results is to build a more sophisticated consumer heterogeneity

Table 3 Prediction Results for Out-of-Sample Data

Method	Log likelihood (LL)	Hit rate
	Consumer homogeneity	
MNL	−4,048.6055	0.512
MEM	−4,048.8774	0.5111
NestL	−4,048.7111	0.5100
	Consumer partworth heterogeneity	
β-HetMNL ($\gamma = 4$)	−3,256.0218	0.6194
β-HetMEM ($\gamma = 4$)	−3,259.6518	0.6180
	Consumer scale heterogeneity MEM	
α-HetMEM ($\gamma = 6$)	−3,201.4232	0.6206
	Consumer scale heterogeneity Nested Logit	
θ-HetNest ($\gamma = 0.7$)	−3,592.2794	0.5491

Note. Results are indicated for the values of γ selected by out-of-sample cross-validation.

Table 4 Results for β-HetMNL Model in $\gamma = 4$ Case

γ	Step	Fit LL	Fit hit rate	Predict LL	Predict hit rate	Time (seconds)
4	1	−4,802.0127	0.7905	−3,652.9236	0.5986	15.699
4	2	−4,092.8787	0.8210	−3,364.9695	0.6254	11.958
4	3	−3,891.1753	0.8178	−3,293.8904	0.6211	11.903
4	4	−3,827.1519	0.8178	−3,276.5331	0.6194	12.623
4	5	−3,789.0618	0.8200	−3,269.1673	0.6174	14.311
4	6	−3,763.5864	0.8227	−3,264.8488	0.6174	16.450
4	7	−3,747.0119	0.8225	−3,261.75611	0.6189	16.515
4	8	−3,736.4616	0.8233	−3,259.3701	0.6174	19.658
4	9	−3,729.7236	0.8232	−3,257.5152	0.6180	36.525
4	10	−3,725.5119	0.8223	−3,256.0216	0.6194	60.321

model that accounts for both partworth and scale heterogeneity as in Fiebig et al. (2010). However, the problem becomes computationally intractable because of the difficulty in calibrating the **D** matrix for this nonconvex problem. To get around such difficulty, we used only the simplest case where **D** was set to the identity matrix and tested the performance of our model, by allowing all or some of the **β** parameters to be heterogeneous. The problem is formulated as follows:

(α, β-HetMEM)

$$\max_{\boldsymbol{\lambda}, \boldsymbol{\beta}_i, \alpha_{ik}, \alpha_1} \left\{ \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} \sum_{k\in\mathcal{K}} z_{ijk} \alpha_{ik} (\boldsymbol{\beta}_i' \mathbf{x}_{ijk} - \lambda_{ij}) - \gamma_1 \sum_{i\in\mathcal{I}} (\alpha_{i1} - \alpha_1)^2 - \gamma_2 \sum_{i\in\mathcal{I}} \|\beta_i - \boldsymbol{\beta}_{i0}\|^2 \right\}$$

$$\text{s.t.} \quad \sum_{k\in\mathcal{K}} e^{\alpha_{ik}(\boldsymbol{\beta}_i'\mathbf{x}_{ijk} - \lambda_{ij})} = 1, \quad i \in \mathcal{I}, j \in \mathcal{J}, \tag{33}$$

$$\sum_{k\in\mathcal{K}} \alpha_{ik} = 4, \quad i \in \mathcal{I},$$

$$\alpha_{i1} = \alpha_{i2} = \alpha_{i3}, \quad i \in \mathcal{I},$$

$$\alpha_{ik} \geq \text{TOL}, \quad i \in \mathcal{I}, k \in \mathcal{K}.$$

To avoid overfitting with many variables, we test the performance again using only a partial set of **β** to depend on consumer demographics—we allow partworths corresponding to the levels of attributes BZ, FP, RP, PP, and NV to differ among different consumers. All other partworth parameters are consumer invariant. From Table 5, we can see that the above modification has only slight improvement in the in-sample and out-of-sample performances. For this data set, modeling alpha and beta heterogeneity together did not provide significant improvement on the performance of α-HetMEM in out-of-sample experiments. This suggests that in this data set it is important to model the scale carefully by allowing for consumer heterogeneity in their perception of the outside option. Using the scale heterogeneity models (α-HetMEM) in this data set provides a significant improvement in terms of out-of-sample performance and better results than partworth heterogeneity models (β-HetMNL or β-HetMEM). Furthermore, it requires only slightly more computational time than the classical MNL model, but improves upon its hit rate performance by more than 20%.

Table 5 **Results for α, β-HetMEM**

γ_1	γ_2	Fit LL	Fit hit rate	Predict LL	Predict hit rate
5	10	−5,311.8948	0.614	−3,198.9593	0.6202
4	10	−5,285.8087	0.6155	−3,212.5870	0.6194
6	10	−5,350.9046	0.615	−3,208.0522	0.62
5	9	−5,300.6860	0.6157	−3,199.0236	0.6203
5	11	−5,321.1340	0.6128	−3,198.9718	0.62

Table 6 **Robustness to Varying Splits of Data Set**

	MNL		α-HetMEM			
Split	Predict LL	Predict hit rate	Predict LL	Predict hit rate	TOL	γ
15:4	−2,304.20266	0.5235	−1,823.1053	0.6315	0.1	4
12:7	−4,048.6055	0.5120	−3,201.4232	0.6206	0.1	6
9:10	−5,850.9678	0.5036	−4,803.7499	0.6000	0.1	5
6:13	−7,723.9843	0.4835	−6,597.5596	0.5746	0.1	4
3:16	−9,800.66147	0.45512	−9,284.8399	0.51487	0.1	30
3:16	−9,800.66147	0.45512	−9,126.9715	0.5220	0.2	8
3:16	−9,800.66147	0.45512	−8,915.5699	0.5246	0.5	4

We finally test for the robustness of the results by splitting the data set in different ratios and compare the homogeneous MNL model with α-HetMEM. Out of the 19 choice tasks, we split the calibration to prediction set sizes in ratios of 15:4, 12:7 (the original split), 9:10, 6:13, and 3:16. The results are provided in Table 6. From the table, we see that the improvements obtained over the simple MNL model are fairly robust, in that the α-HetMEM provides improvement in terms of out-of-sample hit rates from 15% to 20%.

5.3. Standard Error Estimation: Comparison with Mixed Logit

For the second set of experiments, we test the accuracy of the standard error estimation from MDM. For nonparametric methods such as β-HetMNL, such estimates need to be computed using bootstrap methods. For MDM, the optimality conditions and results from §3.5 can be used to compute standard error estimates. We use the MMM class from MDM to test the accuracy of the standard error estimation. Note that we only need to specify the means and variances of the marginal distributions in the MMM model. We compare the performance of MMM with the mixed logit model. Note that mixed logit requires extensive simulation for parameter estimation as well as choice probabilities calculation. We used the *R* software to implement the mixed logit model, and LOQO to solve the MMM.

The demographic attributes for consumer i is denoted by $\mathbf{d}_i$. We model them as continuous variables. We consider three demographic variables in our experiments: (1) income, (2) age, and (3) night (the percentage of time consumer drives during night time), and use the following utility specification for the two models:

$$\tilde{U}_{ik} = (\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}^a)' \mathbf{x}_{ijk} + \boldsymbol{\beta}_d' \mathbf{d}_i + \tilde{\epsilon}_{ik}, \quad i \in \mathcal{I}, j \in \mathcal{J}, k \in \{1, 2, 3\}, \tag{34}$$

$$\tilde{U}_{ij4} = \tilde{\epsilon}_{ij4}, \quad i \in \mathcal{I},$$

where $\tilde{\boldsymbol{\epsilon}}^a$ models the taste variation in **β**. We assume also that the demographic variables affect the utility of the first three alternatives, but not the fourth (cf. Hoffman and Duncan 1988). Under the mixed logit model we assume that individual members of $\tilde{\boldsymbol{\epsilon}}^a$ are i.i.d. normal with zero means, and $\tilde{\epsilon}_{ik}$ are i.i.d. extreme value type-I distributed. For this mixed logit model, $\boldsymbol{\beta}$ and variances of $\tilde{\boldsymbol{\epsilon}}^a$ are estimated by maximizing a simulated likelihood function generated through pseudo random number generations for $\tilde{\boldsymbol{\epsilon}}^a$. Under MMM, mean of utility for each choice is $\boldsymbol{\beta}'\mathbf{x}_{ijk} + \boldsymbol{\beta}_d'\mathbf{d}_i$, and variance is $\mathbf{x}_{ijk}'\Sigma\mathbf{x}_{ijk} + \pi^2/6$. With these, the MMM probabilities can be used for estimation and prediction. We refer to MMM with random coefficients as "mixed MMM."

We used the same setup as before to conduct the second experiment. We allow partworths corresponding to the levels of attributes BZ, FP, RP, PP, and NV to be random, while partworths of other attributes were assumed to be deterministic. We have performed the same experiment on other selections on product attributes, and the results are largely identical. We report the results for this experiment merely because the LL obtained from these models are the best among all experiments performed.

Tables A.1 and A.2 in the appendix show the parameter estimates corresponding to attribute levels, and the associated standard errors, from simulation (for ML) and from the MDM approach (for MMM) based on results in §3.5. A total of 74 parameters were estimated. It is interesting that the mixed MMM can generate maximum likelihood estimators that are fairly similar to the mixed logit, a model that is widely used in practice and well supported in theory. This seems to be in general true for the MLE of means partworths of attribute levels. Furthermore, based on the standard errors calculated, both models identify the same set of partworth attribute levels that are significant. Figures 5 and 6 show the scatter plots of the partworths and the corresponding *p*-values, under both the MMM and MLE models.

Figure 5 **The *p*-Value Estimation**

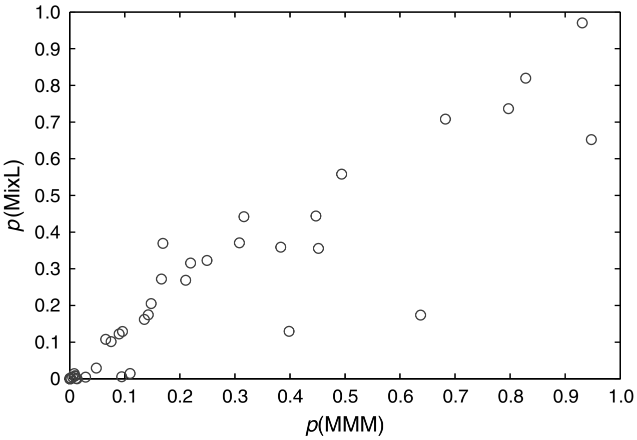
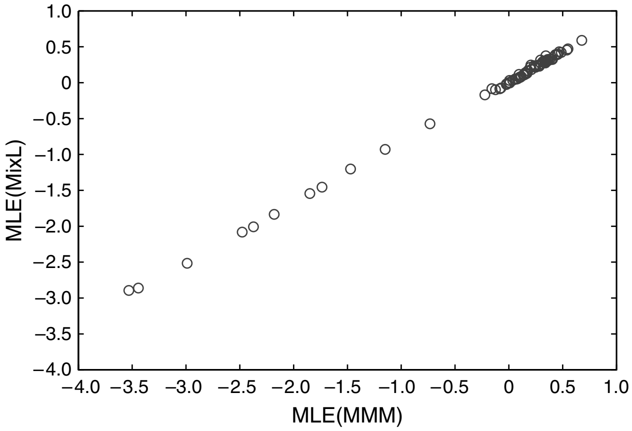


Figure 6 **Partworth Estimates**



5.4. Managerial Insights

Our experiments on the vehicle features data provide some interesting managerial insights. For example, we observe that a technically advanced feature is not always preferred by the consumer over the less advanced one. For the feature attribute navigation system (NS), the levels NS3 and NS4 are technically more advanced as compared with NS1, the former two having extra features on top of the latter. Interestingly, the value of partworth estimates for NS1 is higher than NS3 and NS4. A similar observation holds for attribute lane departure (LD) as well. There can be several reasons for consumers to prefer a less technically advanced product over a technically more advanced one. It may be due to the lack of prior experience with such new technologies, or because consumers prefer human in-the-loop kind of technologies. Another interesting implication of this observation is that a bundle of two attributes is not always desired by consumers. It is possible, for example, that consumers prefer an attribute when presented in the absence of another desired attribute more than a bundle of the two. Our studies also reveal that explicitly showing "none" for the attribute front collision protection (FP) would have a significant negative utility compared to not showing the attribute at all.

The relative preferences across attributes also provide useful insights on willingness to pay for various attribute levels. For example, the MNL as well as the MMM suggest that NV2, NS1, NS4, SC3, and BU4 are the top five attribute levels in terms of highest partworth estimates and consequently highest willingness to pay. Such insights can be used for evaluating proposed feature packages on vehicles and for determining the required pricing for these packages to obtain a balance between demand and profits.

The test implementations of the mixed models suggest there can be significant heterogeneity in preferences of certain attributes like blind zone alert and parallel park aids, and it might be prudent to look at underlying taste variations among the consumers.

6. Conclusion

In this paper, we have provided new theoretical and empirical insights into the semiparametric approach introduced in Natarajan et al. (2009) for discrete choice modeling. Using appropriate choices of marginal distributions, we can reconstruct many interesting choice probability formulas such as MNL, nested logit, and GEV. This allows us to estimate and calibrate the parameters in these models using standard optimization tools, from which the convex nature of many classical log-likelihood maximization models follow easily. The numerical results using a set of conjoint data confirm the efficacy of this approach. By explicitly incorporating consumer heterogeneity into the choice model, we develop a new parsimonious choice model called α-HetMEM and show that it is possible to obtain good out-of-sample performance by allowing for the scale parameter of the outside option to be potentially different across respondents. Capturing scale heterogeneity with a simple model in this data set provides significant improvements over other representations of heterogeneity. In a separate experiment using the same set of data, the mixed form of MMM essentially picks up the same set of significant parameters as the mixed logit model, albeit at a fraction of the computational efforts needed.

The insights obtained in this study are potentially useful for other settings. In most operations management problems involving choice models, invariably the outside option (i.e., "no purchase") is always one of the choices available. This paper proposes a parsimonious and yet computationally tractable method to incorporate this phenomenon into choice prediction, allowing for consumer level heterogeneity in the idiosyncratic error distribution across product options. The performance of this method on an empirical data set is encouraging—the α-HetMEM model retains the simplicity of the MNL method, and yet it can improve the out-of-sample hit rate by close to 20%. A natural extension is to examine the performance of this approach in a revenue management setting, where the MNL has been used predominantly because of its ease of use.

Although we restrict our discussion to discrete choice models, it should be noted that the model of Natarajan et al. (2009) applies to general 0-1 optimization problems. This would allow to extend some of the results in this paper to estimate the partworths of more complicated choice tasks, when there are trade-offs and constraints in the selection of products (such as in the assignment problem, and the shortest path problem). This is potentially useful also for noncompensatory choice modeling, if we impose an additional constraint (with random parameters) into the MDM problem. This can be used to model the ad hoc screening process used by consumers in choosing products to evaluate, and will entail generalizing the MDM model into the case where the objective and constraints are both randomly generated.

Another interesting issue is to understand how the identification for the α and β parameters may be affected by the dimensionality of the problem, measured by the number of alternatives as well as number of attributes in the problem. What is the effect of dimensionality in the quality of the identification? We leave this and other issues for future research.

Acknowledgments

The authors thank Eric Bradlow (special issue coeditor), the associate editor, and two anonymous reviewers for their valuable comments and detailed suggestions on improving this paper. The authors thank the staff from the former General Motors India Laboratory for initiating this project. The research in this paper was funded by the NUS-GM research [Grant R-314-000-087-597]. The research of Vinit Kumar Mishra and Xiaobo Li was also partly funded by the SUTD-MIT International Design Center [Grant IDG31300105] on "Optimization for Complex Discrete Choice."

Appendix

Proof of Theorem 2

Proof. Under MDM, the choice probabilities are given as

$$P_{ik} = 1 - F_{ik}(\lambda_i - V_{ik}) \quad \forall\, k \in \mathcal{K}$$

where λ_i is the solution to the normalization condition:

$$\sum_{k\in\mathcal{K}} (1 - F_{ik}(\lambda_i - V_{ik})) = 1.$$

Under the assumption that $F_{ik}(\cdot)$ is strictly increasing on its support, it is clear that a unique value of the multiplier λ_i satisfies the normalization condition. For the semi-infinite support, this value of λ_i will lie in the range $(\max_{k\in\mathcal{K}}(V_{ik} + \epsilon_{ik}), \infty)$, and for the infinite support, the value of λ will lie in the range $(-\infty, \infty)$. Hence, the choice probabilities strictly satisfy $P_{ik} \in (0, 1)$ with $\sum_k P_{ik} = 1$. To complete the proof, we show that for choice probabilities in the interior of the simplex, there is a unique set of values for the utilities $(V_{i2}, \ldots, V_{iK})$ with $V_{i1} = 0$ and the multiplier λ_i. From the set of equalities, we get

$$\lambda_i = F_{i1}^{-1}(1 - P_{i1})$$

and

$$V_{ik} = F_{i1}^{-1}(1 - P_{i1}) - F_{ik}^{-1}(1 - P_{ik}) \quad \forall\, k \in \mathcal{K}.$$

Hence, there is a one-to-one mapping between the set of deterministic utilities in $\Re_{K-1}$ and the set of choice probabilities in the interior of the unit simplex. □

The exponential marginal distribution for MDM that recreates the MNL choice probabilities, the generalized exponential distribution for MDM that recreates GEV choice probabilities, and the t-distribution for MDM that recreates the MMM choice probabilities satisfy the conditions in Theorem 2.

Identifiability of V_i and F_{ik}

Assume that V_{ik} is an affine function of the attributes of the kth alternative, i.e., $V_{ik} = \beta'\mathbf{x}_{ik}$. Let $\mathcal{F}$ be a prespecified family of distributions containing possible error distributions. We would like to answer the following question:

Are there distinct solutions $(\beta^a, \{F_{ik}^a(\cdot)\})$ and $(\beta^b, \{F_{ik}^b(\cdot)\})$, with $F_{ik}^a(\cdot), F_{ik}^b(\cdot) \in \mathscr{F}$, such that

$$1 - F_{ik}^a(\lambda_i^a(\mathbf{x}) - (\beta^a)'\mathbf{x}_{ik}) = 1 - F_{ik}^b(\lambda_i^b(\mathbf{x}) - (\beta^b)'\mathbf{x}_{ik}) \text{ for all } k,$$

and

$$\sum_{k\in\mathscr{K}}(1 - F_{ik}^a(\lambda_i^a(\mathbf{x}) - (\beta^a)'\mathbf{x}_{ik})) = \sum_{k\in\mathscr{K}}(1 - F_{ik}^b(\lambda_i^b(\mathbf{x}) - (\beta^b)'\mathbf{x}_{ik})) = 1,$$

for some $\lambda_i^a(\mathbf{x})$, $\lambda_i^b(\mathbf{x})$, given $\mathbf{x}_{ik}$?

If we fix $(\beta^a)'\mathbf{x}_{i1} = (\beta^b)'\mathbf{x}_{i1} = 0$ and standardize $F_{i1}^a(\cdot) = F_{i1}^b(\cdot)$, then we have $1 - F_{i1}^a(\lambda_i^a(\mathbf{x})) = 1 - F_{i1}^b(\lambda_i^b(\mathbf{x}))$, and hence $\lambda_i^a(\mathbf{x}) = \lambda_i^b(\mathbf{x})$. If the family $\mathscr{F}$ is chosen in such a way that $F_{ik}^a(\lambda_i(\mathbf{x}) - (\beta^a)'\mathbf{x}_{ik}) = F_{ik}^b(\lambda_i(\mathbf{x}) - (\beta^b)'\mathbf{x}_{ik})$ for all $\mathbf{x}_{ik}$ only if $F_{ik}^a(\cdot) = F_{ik}^b(\cdot)$ and $\beta^a = \beta^b$, then we would be able to identify β and $F_{ik}(\cdot)$ from the choice probabilities. For instance, when $\mathscr{F} = \{F_{ik}(x) = 1 - e^{-\alpha_{ik}x}\colon \alpha_{ijk} > 0\}$, then $F_{ik}^a(\lambda_i(\mathbf{x}) - (\beta^a)'\mathbf{x}_{ik}) = F_{ik}^b(\lambda_i(\mathbf{x}) - (\beta^b)'\mathbf{x}_{ik})$ holds if and only if $\alpha_{ik}^a(\lambda_i(\mathbf{x}) - (\beta^a)'\mathbf{x}_{ik}) = \alpha_{ik}^b(\lambda_i(\mathbf{x}) - (\beta^b)'\mathbf{x}_{ik})$ for all $\mathbf{x}_{ik}$. Since the Lagrange multiplier $\lambda_i(\mathbf{x})$ cannot be a linear function of $\mathbf{x}$, this is possible only if $\alpha_{ik}^a = \alpha_{ik}^b$ and $\beta^a = \beta^b$.

Proof of Theorem 4

We rewrite problem (15) in the following manner.

$$\begin{aligned}\max_{\lambda,\beta}\ &\Bigg\{\sum_{i\in\mathscr{I}}\sum_{k\in\mathscr{K}} z_{ik}\Bigg(-\lambda_i + \beta'\mathbf{x}_{ik}\\ &\qquad + (\theta_r - 1)\ln\Bigg(\sum_{l\in B_r(k)} e^{(1/\theta_r)(\beta'\mathbf{x}_{\mathrm{il}} - \beta'\mathbf{x}_{\mathrm{ik}})}\Bigg)\Bigg)\Bigg\} \qquad (35)\\ \text{s.t. } &\sum_{r=1}^{R}\Bigg(\sum_{l\in B_r} e^{(\beta'\mathbf{x}_{\mathrm{il}})/\theta_r}\Bigg)^{\theta_r} \times e^{-\lambda_i} \le 1, \quad i\in\mathscr{I}.\end{aligned}$$

Since $\theta \in (0,1]$ and $\ln\sum_i e^{x_i}$ is convex in $\mathbf{x}$, the objective function of this problem is concave. To show that the feasible region is convex, it suffices to show that the function $g(\mathbf{x}) = (\sum_{i\in K} e^{x_i/a})^a$ is convex in $\mathbf{x}$ for $a \in (0,1]$. The Hessian matrix $\nabla^2(g(\mathbf{x}))$ of this function is $((\mathbf{1}'\mathbf{z})^{a-2}/a)((a-1)\mathbf{z}\mathbf{z}' + (\mathbf{1}'\mathbf{z})\mathrm{diag}(\mathbf{z}))$, where $\mathbf{z} = (e^{x_1/a}, e^{x_2/a}, \ldots)$, $\mathrm{diag}(\mathbf{z})$ is the diagonal matrix with elements of $\mathbf{z}$ in diagonal, and $\mathbf{1}$ is the vector of ones. Note that $(\mathbf{1}'\mathbf{z})^{a-2}/a > 0$ for $a \in (0,1]$. Furthermore,

$$((a-1)\mathbf{z}\mathbf{z}' + (\mathbf{1}'\mathbf{z})\mathrm{diag}(\mathbf{z})) = (a\mathbf{z}\mathbf{z}' + \mathbf{1}'\mathbf{z}\ \mathrm{diag}(\mathbf{z}) - \mathbf{z}\mathbf{z}').$$

Clearly, $\mathbf{z}\mathbf{z}' \succeq 0$. Furthermore, $\mathbf{1}'\mathbf{z}\ \mathrm{diag}(\mathbf{z}) - \mathbf{z}\mathbf{z}' \succeq 0$ due to the Cauchy–Schwarz inequality. The convexity of $g(\mathbf{x})$ follows from the positive-semidefiniteness of Hessian $\nabla^2(g(\mathbf{x}))$. □

Heterogeneity in MNL Type Models

Note that there are several ways in which one could introduce heterogeneity into the scale parameters in MNL type models. If we simply allow each β_i to be $\alpha_i\beta$ for some value α_i and fixed β, and add a regularization term as before to penalize deviation of α_i from some number α_0, then the calibration problem indeed becomes much simpler than β-HetMNL. The model is

$$\begin{aligned}(\alpha\text{-HetMNL})\quad \max_{\lambda,\beta,\alpha_i,\alpha_0}\ &\Bigg\{\sum_{i\in\mathscr{I}}\sum_{j\in\mathscr{J}}\sum_{k\in\mathscr{K}} z_{ijk}\ln\Bigg(\frac{e^{\alpha_i\beta'\mathbf{x}_{ijk}}}{\sum_{l\in\mathscr{K}} e^{\alpha_i\beta'\mathbf{x}_{ijl}}}\Bigg)\\ &\quad -\gamma\sum_{i\in\mathscr{I}}(\alpha_i - \alpha_0)^2\Bigg\} \qquad (36)\\ \text{s.t. } &\mathrm{DOWN} \le \alpha_i \le \mathrm{UP}, \quad i\in\mathscr{I}.\end{aligned}$$

This model is indeed similar to α-HetMEM, except that its emphasis is on the scale of every consumer, whereas in α-HetMEM, the emphasis is on the perception of the consumer in the outside option. The results are provided in the following table.

DOWN	UP	γ	Fit LL	Fit HR	Predict LL	Predict HR
0.1	4	0	−6,253.1878	0.4561	−3,621.6659	0.5082
0.1	4	0.5	−6,328.4420	0.462	−3,617.4705	0.5128
0.1	4	1	−6,360.9836	0.465	−3,622.7267	0.5148
0.1	2	0	−6,305.7746	0.4597	−3,617.2259	0.5126
0.1	2	0.5	−6,328.5502	0.462	−3,617.5081	0.5128
0.1	2	1	−6,360.9836	0.465	−3,622.7266	0.5148
1	2	0	−6,816.7917	0.4706	−3,822.9386	0.5094
1	2	0.5	−6,819.5745	0.4705	−3,822.8581	0.5097
1	2	1	−6,828.4274	0.4706	−3,827.5302	0.5097

By varying the values of γ, UP, and DOWN to test the performance, the results show that the prediction performance (LL $\approx -3{,}620$) lies in the halfway between MNL (LL $= -4{,}048.6055$) and α-HetMEM (LL $= -3{,}201.4232$). Hence, this version of α-HetMNL does not perform as well.

An alternative approach is to use the HEV model with Gumbel random variables with different scales. However, since the HEV model does not possess closed-form choice probabilities, it is computationally challenging to try and estimate the heteroskedastic parameter for every consumer with regularization terms. The simulated experiments in §4 indicate that the α-HetMEM is fairly good at capturing the parameters coming from the HEV model. The last model that we experimented with was a modified MNL (MMNL) in the spirit of our α-hetMEM. The calibration under this problem is

$$\begin{aligned}(\alpha\text{-HetMMNL})\quad \max_{\lambda,\beta,\alpha_{ik},\alpha_0}\ &\Bigg\{\sum_{i\in\mathscr{I}}\sum_{j\in\mathscr{J}}\sum_{k\in\mathscr{K}} z_{ijk}\ln\Bigg(\frac{e^{\alpha_{ik}\beta'\mathbf{x}_{ijk}}}{\sum_{l\in\mathscr{K}} e^{\alpha_{il}\beta'\mathbf{x}_{ijl}}}\Bigg)\\ &\quad -\gamma\sum_{i\in\mathscr{I}}(\alpha_{i1} - \alpha_1)^2\Bigg\} \qquad (37)\\ \text{s.t. } &\sum_{k\in\mathscr{K}}\alpha_{ik} = 4, \quad i\in\mathscr{I},\\ &\alpha_{i1} = \alpha_{i2} = \alpha_{i3}, \quad i\in\mathscr{I},\\ &\alpha_{ik} \ge \mathrm{TOL}, \quad i\in\mathscr{I}, k\in\mathscr{K}.\end{aligned}$$

Although this model has similarities to the α-HetMEM model, the performance is not as good. The following table provides the results:

γ	Fit LL	Fit HR	Predict LL	Predict HR
0.1	−6,350.2662	0.4615	−3,623.0763	0.514
0.5	−6,354.1941	0.462	−3,620.1574	0.5134
2	−6,399.0739	0.4661	−3,632.3006	0.5134

Again the results indicate that both the in-sample and out-of-sample performances are similar to the α-HetMNL model, which is not as good as α-HetMEM.

Comparison of Mixed MMM and Mixed Logit

Table A.1 Estimation Results for Mixed MMM and Mixed Logit Model-I

	Mixed MMM					Mixed logit			
Parameter	MLE	Std. error	t-value	Pr $>$ $\lvert t\rvert$	Parameter	MLE	Std. error	t-value	Pr $>$ $\lvert t\rvert$
cc1***	0.398662	0.09054572	4.40288069	1.08705E−05	cc1***	0.32574	0.076165	4.2768	0.00001896
cc2***	0.339607	0.088727307	3.827536451	0.000130769	cc2***	0.27439	0.074179	3.699	0.0002164
cc3***	0.36051	0.089652707	4.021183641	5.86277E−05	cc3***	0.30695	0.074385	4.1265	0.00003683
gn1	0.115812	0.079975188	1.448099121	0.147642378	gn1	0.086939	0.068643	1.2665	0.2053253
gn2	0.0803683	0.080173937	1.002424269	0.316179766	gn2	0.052636	0.068456	0.7689	0.4419485
ns1***	0.540169	0.102108237	5.290160884	1.26607E−07	ns1***	0.45493	0.084169	5.405	6.48E−08
ns2	0.0358746	0.103254123	0.34743988	0.728273249	ns2	0.032526	0.086865	0.3744	0.708079
ns3	0.151385	0.106853961	1.416746734	0.156609696	ns3	0.13628	0.088261	1.5441	0.1225607
ns4***	0.487241	0.103093018	4.726226927	2.3403E−06	ns4***	0.42011	0.083152	5.0523	4.364E−07
ns5	−0.124014	0.104530081	−1.186395326	0.235513907	ns5	−0.098734	0.089872	−1.0986	0.2719379
bu1*	0.247324	0.109618994	2.256214822	0.024093575	bu1*	0.22245	0.090652	2.4539	0.0141332
bu2**	0.315278	0.107954458	2.920472248	0.00350823	bu2**	0.28619	0.090938	3.1471	0.0016488
bu3***	0.432823	0.109196941	3.963691632	7.46698E−05	bu3***	0.38938	0.090135	4.3199	0.00001561
bu4***	0.432593	0.112279847	3.852810733	0.000118006	bu4***	0.38393	0.089911	4.2701	0.00001954
bu5**	0.314371	0.112985732	2.782395571	0.005413044	bu5**	0.29076	0.091585	3.1748	0.0014993
bu6	−0.0833997	0.112864517	−0.738936399	0.459974964	bu6	−0.085001	0.092655	−0.9174	0.3589369
fa1***	0.279762	0.080407507	3.479301999	0.000506349	fa1***	0.24716	0.068751	3.595	0.0003243
fa2	0.0608508	0.080027742	0.760371324	0.447062992	fa2	0.052879	0.069081	0.7655	0.4439982
ld1***	0.361308	0.089857016	4.02092143	5.86928E−05	ld1***	0.32315	0.075978	4.2532	0.00002108
ld2**	0.283853	0.0882024	3.218200412	0.001296941	ld2**	0.2294	0.073996	3.1002	0.0019341
ld3	0.133014	0.089058556	1.493556662	0.135344917	ld3	0.10749	0.076837	1.3989	0.1618506
bz1*	0.29458	0.118661916	2.482515115	0.013073408	bz1***	0.31615	0.075622	4.1806	0.00002907
bz2**	0.344085	0.109854458	3.132189694	0.001743554	bz2***	0.3754	0.078809	4.7635	0.000001903
bz3	−0.0233921	0.107942655	−0.216708586	0.828442936	bz3	−0.01938	0.084974	−0.2281	0.819593
fc1***	0.367185	0.079603264	4.612687736	4.05878E−06	fc1***	0.31237	0.068204	4.5799	0.000004651
fc2	0.101607	0.081161943	1.25190448	0.210654216	fc2	0.077122	0.06974	1.1059	0.2687908
fp1	0.204973	0.122518399	1.672997699	0.094380675	fp1**	0.24487	0.08885	2.756	0.0058511
fp2***	0.448341	0.11362379	3.945837389	8.04433E−05	fp2***	0.38893	0.081675	4.7619	0.000001917
fp3	0.199766	0.124837987	1.600202033	0.109607222	fp3*	0.21386	0.087592	2.4416	0.0146223
fp4	−0.223402	0.121143679	−1.844107772	0.065217457	fp4	−0.17036	0.10591	−1.6085	0.1077203
rp1*	0.235401	0.093065203	2.529420163	0.011450898	rp1***	0.23578	0.070084	3.3642	0.0007676
rp2	0.00658772	0.100388209	0.065622448	0.947680636	rp2	0.031849	0.070654	0.4508	0.6521464

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Table A.2 Estimation Results for Mixed MMM and Mixed Logit Model-II

	Mixed MMM					Mixed logit			
Parameter	MLE	Std. error	t-value	Pr $>$ $\lvert t\rvert$	Parameter	MLE	Std. error	t-value	Pr $>$ $\lvert t\rvert$
pp1***	0.405399	0.106963117	3.790082148	0.000152096	pp1***	0.32397	0.078736	4.1147	0.00003878
pp2*	0.234565	0.107242461	2.187239996	0.028764062	pp2**	0.23175	0.081958	2.8276	0.0046897
pp3	0.0948428	0.112283625	0.844671693	0.398328359	pp3	0.11533	0.076031	1.5169	0.1292989
ka1***	0.333277	0.080610221	4.134426055	3.6077E−05	ka1***	0.29162	0.070169	4.156	0.00003239
ka2**	0.209422	0.080676056	2.595838351	0.009459391	ka2**	0.18243	0.069632	2.6199	0.0087961
sc1***	0.339823	0.095637405	3.55324363	0.000383482	sc1***	0.28495	0.080582	3.5362	0.000406
sc2**	0.301116	0.096262222	3.1280807	0.001768074	sc2**	0.26195	0.07968	3.2875	0.0010108
sc3***	0.549228	0.096989831	5.662737987	1.55991E−08	sc3***	0.47159	0.08033	5.8707	4.34E−09
sc4	0.00829448	0.095891568	0.086498533	0.931073052	sc4	−0.0030269	0.081662	−0.0371	0.9704321
ts1***	0.3833	0.089413697	4.286815273	1.84136E−05	ts1***	0.32855	0.075344	4.3606	0.00001297
ts2***	0.457622	0.089413218	5.1180576	3.18456E−07	ts2***	0.39905	0.075332	5.2972	1.176E−07
ts3*	0.177221	0.089631421	1.977219572	0.048063213	ts3*	0.16714	0.076673	2.1799	0.0292661
nv1***	0.467919	0.105160324	4.44957737	8.76185E−06	nv1***	0.42944	0.077168	5.565	2.621E−08
nv2***	0.677672	0.104949418	6.457129644	1.15211E−10	nv2***	0.5896	0.078425	7.5179	5.573E−14
nv3	−0.157573	0.114562655	−1.375430765	0.169050066	nv3	−0.083872	0.093388	−0.8981	0.3691336
ma1	0.159165	0.09554434	1.665875755	0.09579116	ma1	0.12147	0.080026	1.5179	0.1290386
ma2	0.0664014	0.097047423	0.684216004	0.493865605	ma2	0.047416	0.080955	0.5857	0.5580738
ma3***	0.34282	0.096013298	3.57054709	0.000359068	ma3***	0.28601	0.079795	3.5843	0.000338

Table A.2 (Continued)

	Mixed MMM					Mixed logit			
Parameter	MLE	Std. error	*t*-value	Pr > |*t*|	Parameter	MLE	Std. error	*t*-value	Pr > |*t*|
ma4	0.110704	0.096065334	1.152382398	0.249210581	ma4	0.079377	0.080246	0.9892	0.3225843
lb1	0.0965966	0.094774553	1.019225061	0.308137852	lb1	0.070988	0.07929	0.8953	0.3706311
lb2	0.143061	0.097531122	1.466824095	0.142477121	lb2	0.10881	0.080141	1.3578	0.1745426
lb3	0.169312	0.0950662	1.780990502	0.07496535	lb3	0.13181	0.080465	1.6382	0.1013889
lb4	−0.0722387	0.095989094	−0.752571955	0.451737149	lb4	−0.074834	0.080996	−0.9239	0.3555237
af1***	0.365297	0.087662821	4.167068719	3.12944E−05	af1***	0.30735	0.074125	4.1464	0.00003377
af2**	0.247988	0.088843965	2.791275682	0.005266925	af2**	0.21493	0.074818	2.8728	0.0040691
af3	−0.0235715	0.091693143	−0.257069386	0.797134159	af3	−0.026228	0.077915	−0.3366	0.7364008
hu1***	0.349939	0.079554898	4.398710952	1.10808E−05	hu1***	0.29532	0.068467	4.3133	0.00001609
hu2	0.0993129	0.080882392	1.227867995	0.219545426	hu2	0.07005	0.06983	1.0032	0.3157862
Price2***	−0.732407	0.130178037	−5.626194849	1.92683E−08	Price2***	−0.57239	0.083242	−6.8762	6.147E−12
Price3***	−1.1495	0.131429762	−8.746116392	2.84051E−18	Price3***	−0.92979	0.087734	−10.5978	< 2.2e − 16
Price4***	−1.47071	0.135820887	−10.8283051	4.53197E−27	Price4***	−1.203	0.093321	−12.8908	< 2.2e − 16
Price5***	−1.73634	0.138530497	−12.53399097	1.37502E−35	Price5***	−1.456	0.097447	−14.9411	< 2.2e − 16
Price6***	−1.84914	0.140580221	−13.15362851	5.69824E−39	Price6***	−1.5448	0.099941	−15.4569	< 2.2e − 16
Price7***	−2.18123	0.151474296	−14.40000093	3.11346E−46	Price7***	−1.8345	0.10464	−17.5316	< 2.2e − 16
Price8***	−2.37265	0.153351597	−15.47196149	5.79821E−53	Price8***	−2.0085	0.10753	−18.6783	< 2.2e − 16
Price9***	−2.47772	0.161114507	−15.37862761	2.32731E−52	Price9***	−2.0826	0.11033	−18.8756	< 2.2e − 16
Price10***	−2.98843	0.184297331	−16.21526465	6.88454E−58	Price10***	−2.5157	0.12562	−20.0261	< 2.2e − 16
Price11***	−3.53154	0.21556974	−16.38235495	5.03245E−59	Price11***	−2.8943	0.13505	−21.4314	< 2.2e − 16
Price12***	−3.44198	0.20769941	−16.57192957	2.51211E−60	Price12***	−2.8617	0.13181	−21.7101	< 2.2e − 16
agea***	−0.01825794	0.001024016	−17.82974051	2.64459E−69	agea***	−0.016952	0.0023411	−7.2412	4.448E−13
nighta	0.002489172	0.005281256	0.471321934	0.637428252	nighta	0.0023613	0.0017368	1.3595	0.1739863
incomea_n**	1.85482E−06	7.08457E−07	2.61811118	0.008864352	incomea_n**	1.6895E−06	6.2719E−07	2.6938	0.007065

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

References

Allenby GM, Rossi PE (1999) Marketing models of consumer heterogeneity. *J. Econometrics* 89(1–2):57–78.

Anderson SP, De Palma A, Thisse J-F (1988) A representative consumer theory of the logit model. *Internat. Econom. Rev.* 29(3):461–466.

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

Bertsimas D, Natarajan K, Teo CP (2006) Persistence in discrete optimization under data uncertainty. *Math. Programming, Ser. B* 108(2–3):251–274.

Bhat CR (1995) A heteroscedastic extreme value model of intercity mode choice. *Transportation Res. Part B* 29(6):471–483.

Brownstone D, Small KAK (1989) Efficient estimation of nested logit models. *J. Bus. Econom. Statist., Amer. Statist. Assoc.* 7(1):67–74.

Chrzan K, Elrod T (1995) Choice-based approach for large numbers of attributes. *Marketing News* 29(1):20–24.

Daganzo CF, Kusnic M (1993) Two properties of the nested logit model. *Transportation Sci.* 27(4):395–400.

Danaher PJ (2007) Modeling page views across multiple websites with an application to Internet reach and frequency prediction. *Marketing Sci.* 26(3):422–437.

Danaher PJ, Smith MS (2011) Modeling multivaraite distributions using copulas: Applications in marketing. *Marketing Sci.* 30(1):4–21.

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

Farias VF, Jagabathula S, Shah D (2013) A nonparametric approach to modeling choice with limited data. *Management Sci.* 59(2): 305–322.

Fiebig DG, Keane MP, Louviere J, Wasi N (2010) The generalized multinomial logit model: Accounting for scale and coefficient heterogeneity. *Marketing Sci.* 29(3):393–421.

Greene WH (2011) *Econometric Analysis*, 7th ed. (Prentice Hall, Upper Saddle River, NJ).

Hofbauer J, Sandholm WH (2002) On the global convergence of stochastic fictitious play. *Econometrica* 70(6):2265–2294.

Hoffman SD, Duncan GJ (1988) Multinomial and conditional logit discrete-choice models in demography. *Demography* 25(3): 415–427.

Louviere J (1988) Conjoint analysis modelling of stated preferences. A review of theory, methods, recent developments and external validity. *J. Transport Econom. Policy* 22(1):93–119.

Louviere J, Meyer RJ (2007) Formal choice models of informal choices: What choice modeling research can (and can't) learn from behavioral theory. Malhotra NK, ed. *Review of Marketing Research* (M. E. Sharpe, New York), 3–32.

Louviere J, Swait J (2010) Discussion of "alleviating the constant stochastic variance assumption in decision research: Theory, measurement and experimental test." *Marketing Sci.* 29(1): 18–22.

Louviere J, Street S, Carson R, Ainslie A, Deshazo JR, Cameron T, Hensher D, Kohn R, Marley T (2002) Dissecting the random component of utility. *Marketing Lett.* 13(3):177–193.

Manski CF (1975) Maximum score estimation of the stochastic utility model of choice. Zarembka P, ed. *J. Econometrics* 3:205–228.

McFadden D (1974) Conditional logit analysis of qualitative choice behavior. Zarembka P, ed. *Frontiers in Econometrics* (Academic Press, New York), 105–142.

McFadden D (1977) A closed-form multinomial choice model without the independence from irrelevant alternatives restrictions. Working Paper 7703. Urban Travel Demand Forecasting Project, Institute of Transportation Studies, University of California, Berkeley, Berkeley.

McFadden D (1978) Modelling the choice of residential location. Karlqvist A, Lundqvist L, Snickars F, Weibull J, eds. *Spatial Interaction Theory and Planning Models* (North Holland, Amsterdam), 75–96.

Natarajan K, Song M, Teo C-P (2009) Persistency model and its applications in choice modeling. *Management Sci.* 55(3):453–469.

Nelsen RB (2006) *An Introduction to Copulas*, 2nd ed., Springer Series in Statistics (Springer, New York).

Norets A, Takahashi S (2013) On the surjectivity of the mapping betweeen utilities and choice probabilities. *Quant. Econom.* 4(1): 149–155.

Park Y-H, Fader PS (2004) Modeling browsing behavior at multiple websites. *Marketing Sci.* 23(2):280–303.

Salisbury LC, Feinberg FM (2010) Alleviating the constant stochastic variance assumption in decision research: Theory, measurement and experimental test. *Marketing Sci.* 29(1):1–17.

Sawtooth Software (2008) CBC v6.0. Sawtooth Software Inc., Sequim, WA. http://www.sawtoothsoftware.com/download/techpap/cbctech.pdf.

Schweidel DA, Fader PS, Bradlow ET (2008) A bivariate timing model of consumer acquisition and retention. *Marketing Sci.* 27(5):829–843.

Seetharman PB, Chib S, Ainslie A, Boatwright P, Chan T, Gupta S, Mehta N, Rao V, Strijnev A (2005) Models of multi-category choice behavior. *Marketing Lett.* 16(3/4):239–254.

Sklar A (1959) Fonctions de répartition á n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* 8:229–231.

Steenburgh TJ, Ainslie AS (2010) Substitution patterns of the ranndom coefficients logit. HBS Marketing Unit Working Paper 10-053, Harvard Business School, Boston. http://ssrn.com/abstract=1535329.

Train K (2009) *Discrete Choice Methods with Simulation*, Second ed. (Cambridge University Press, Cambridge, UK).

Weiss G (1986) Stochastic bounds on distributions of optimal value functions with applications to PERT, network flows and reliability. *Oper. Res.* 34(4):595–605.