Journal of Choice Modelling 46 (2023) 100399

Contents lists available at ScienceDirect

Journal of Choice Modelling

journal homepage: www.elsevier.com/locate/jocm

A control-function correction for endogeneity in random coefficients models: The case of choice-based recommender systems

Mazen Danaf [a,*], C. Angelo Guevara [b,c], Moshe Ben-Akiva [d]

[a] Uber Freight, 433 W Van Buren St, Chicago, IL, 60607, USA
[b] Departamento de Ingeniería Civil, Universidad de Chile, Blanco Encalada 2002, Santiago, Chile
[c] Instituto Sistemas Complejos de Ingeniería (ISCI), República 695, Santiago, Chile
[d] Edmund K. Turner Professor of Civil and Environmental Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA, 02139, USA

ARTICLE INFO

Keywords:
Discrete choice models
Endogeneity
Control-function
Recommender systems
Logit mixture
Random parameters

ABSTRACT

Applications of discrete choice models in personalization are becoming increasingly popular among researchers and practitioners. However, in such systems, when users are presented with successive menus (or choice situations), the alternatives and attributes in each menu depend on the choices made by the user in the previous menus. This gives rise to endogeneity which can result in inconsistent estimates. Our companion paper, Danaf et al. (2020), showed that the estimates are only consistent when the entire choice history of each user is included in estimation. However, this might not be feasible because of computational constraints or data availability. In this paper, we present a control-function (CF) correction for the cases where the choice history cannot be included in estimation. Our method uses the attributes of **non-personalized** attributes as instruments, and applies the CF correction by including interactions between the explanatory variables and the first stage residuals. Estimation can be done either sequentially or simultaneously, however, the latter is more efficient (if the model reflects the true data generating process). This method is able to recover the population means of the distributed coefficients, especially with a long choice history. The variances are underestimated, because part of the inter-consumer variability is explained by the residuals, which are included in the systematic utility. However, the population variances can be computed from the estimation results. The modified utility equations (which include the residuals) can be used in forecasting and model application, and provide superior fit and predictions.

1. Introduction

Applications of discrete choice models in personalization are becoming increasingly popular among researchers and practitioners (Chaptini, 2005; Danaf et al., 2019; Song et al., 2017, 2018; Teo et al., 2016; Zhu et al., 2019; etc.). These models overcome several limitations associated with traditional recommendation and personalization methods. According to Chaptini (2005), discrete choice models can account for all the available data including user-specific, item-specific, and contextual information (while some traditional

* Corresponding author.
E-mail addresses: mazendanaf@gmail.com (M. Danaf), crguevar@ing.uchile.cl (C.A. Guevara), mba@mit.edu (M. Ben-Akiva).

https://doi.org/10.1016/j.jocm.2022.100399
Received 29 January 2022; Received in revised form 23 December 2022; Accepted 30 December 2022
Available online 31 December 2022

methods rely mainly on item and user profiling). In addition, they express utility as a function of attributes, which enables us to recommend new items that have not been chosen or rated before. Finally, while other personalization methods struggle with finding a balance between relevance and diversity, choice models integrate both (relevance and diversity) in a unified utility-based framework (Jiang et al., 2014).

Many of the recent applications of choice models in recommender systems rely on the logit mixture model, primarily because it can explicitly account for unobserved inter-consumer heterogeneity. By conditioning on the observed choices for each user, we can extract user-specific parameters, which can be used in personalization (Revelt and Train, 2000). However, our companion paper Danaf et al. (2020) showed that estimating discrete choice models in adaptive contexts (such as recommender systems and adaptive stated preferences (ASP) surveys) can - in some cases - result in substantial biases due to endogeneity. In order to obtain consistent estimates of the population parameters, we recommended including the entire choice history of each user in estimation. While this might be feasible in some applications involving small datasets (such as ASP surveys), choice-based recommender systems can be limited by data availability and computational constraints. In this paper, we propose a control-function (CF) correction for addressing endogeneity in such cases.

Endogeneity arises in discrete choice models due to several factors including measurement errors, selection bias, omitted variables, and simultaneity, and results in inconsistent estimates of the model parameters (Guevara, 2015). The textbook definition of endogeneity is a correlation between the independent/observed variables in the model and the unobserved additive error term. In the presence of taste heterogeneity, endogeneity can also be attributed to correlations between the explanatory variables and the heterogeneity (i.e., the individual-specific parameters) (Wooldridge, J.M., 2015).

Several corrections have been proposed for endogeneity in discrete choice models, mainly falling into two categories; the BLP method (Berry et al., 1995), and the CF method (Heckman, 1978; Hausman, 1978). CF methods use extra variables in the utility specification that are obtained using exogenous instruments. These instruments are variables that must exhibit two characteristics. First, they should be *relevant*, meaning that they are correlated with the endogenous variable. Second, they should be *exogenous*, which means that they should not be correlated with the unobserved error term.

Different control-functions have been proposed, the most common of which are by Petrin and Train (2010), Villas-Boas and Winer (1999), Blundell and Powell (2003), Guevara and Ben-Akiva (2006, 2010), and Guevara and Polanco (2016). These methods are convenient because they are easier in implementation, and because they apply to cases where BLP cannot be used (Train, 2009). Most of the proposed CF methods address the issue of correlation between the explanatory variables and the unobserved additive error term. On the other hand, very limited research has addressed the correlation between the explanatory variables and the unobserved heterogeneity, which applies to endogeneity under taste variation (e.g., choice-based recommender systems).

Common examples of endogeneity under taste variation include adaptive stated preferences (ASP) designs, choice-based self-selection, personalized advertisement, and recommender systems. In ASP surveys, the design is updated during the data collection phase as knowledge of the true parameters increases (Kanninen, 2002; Johnson et al., 2006, 2013). Bradley and Daly (1993) analyzed several variations of ASP designs and concluded that endogenous SP designs might result in bias in the presence of taste variation in the sample. Similarly, Fowkes (2007) analyzed endogeneity bias in the Leeds Adaptive Stated Preferences (LASP) survey (Fowkes and Shinghal, 2002; Shinghal, 1999), and found significant bias when the models are calibrated over several respondents. Toubia et al. (2003) and Abernethy et al. (2007), also developed polyhedral methods for survey designs that “reduce the feasible set of parameters as rapidly as possible” but concluded that these methods are susceptible to endogeneity bias.

SP designs that are constructed based on the revealed preferences (RP) choices can also be susceptible to endogeneity bias if estimation is done using the SP data only. Train and Wilson (2008, 2009) propose a full information maximum likelihood (FIML) solution, which overcomes this problem by using a specially designed maximum simulated likelihood method. On the other hand, Guevara and Hess (2019) propose a limited information maximum likelihood (LIML) approach based on a control-function that uses RP attributes as instruments. However, both cases consider correlations between the explanatory variables and the unobserved error term, and not between the explanatory variables and the individual-specific effects.

In this paper, we extend the CF method proposed by Wooldridge (2015) and Garen (1984), and apply it to the case of choice-based personalization methods. Our CF method uses the attributes of *non-personalized* recommendations (i.e., what we would recommend or advertise to an average individual) as instruments to address cases where historical data are excluded from estimation. We apply this method to a Monte Carlo dataset mimicking a choice-based recommender system that recommends the “best” alternatives to users based on their estimated individual-specific preferences, and a real dataset that is modified to introduce endogeneity resulting from personalized advertising. Finally, we extend the analysis of Guevara and Ben-Akiva (2012) to illustrate how the residuals obtained from the first stage estimation can be used in forecasting and in model application.

Although this paper is primarily motivated by choice-based personalization, the proposed CF correction applies to various applications associated with panel data. Such applications are usually encountered in dynamic choice contexts, where serial correlation (e.g., preference heterogeneity) and the inter-dependence of successive choices give rise to the ***initial conditions problem***. In such cases, model estimation under the assumption that the initial conditions or the relevant pre-sample history are exogenous might result in biased and inconsistent parameter estimates. Examples are common in transportation, including cases where individuals with a higher time sensitivity tend to relocate closer to their workplace (resulting in shorter travel times), or where individuals with a higher preference towards transit purchase a transit pass (resulting in a lower cost).

The remainder of this paper is organized as follows. Section 2 presents a brief background on endogeneity in recommender systems. Section 3 presents the proposed CF methodology and its application in forecasting. This methodology is applied to simulated and real data in Section 4. Section 5 presents a discussion on estimation considerations, consistency, and simultaneous estimation. Finally, Section 6 concludes the paper by discussing the contributions, limitations, and future research directions.

2. Background: endogeneity in recommender systems

2.1. Problem statement

Recommender systems are software tools and techniques providing suggestions for items to be of use to a user (Ricci et al., 2015). These techniques are used to cope with information overload and to identify a subset of items from a much larger set (*the inventory*) that best matches a user's interest (Chaptini, 2005; Jiang et al., 2014). In choice-based recommender systems, recommendations are usually generated using individual-specific coefficients that are fed into an assortment optimization, which maximizes hit-rate or consumer-surplus (CS) in the form of log-sum (Song et al., 2017, 2018).

Danaf et al. (2020) analyzed endogeneity bias in choice-based recommender systems using a Monte Carlo experiment and found that the estimates are only consistent when the likelihood function accounts for the full data generation process. This can be achieved by using exogenous initialization of the system and using all the data in estimation. This section provides a brief theoretical analysis of endogeneity in adaptive choice contexts and explains why excluding historical data from estimation results in inconsistency. For more details, we refer the reader to Danaf et al. (2020).

We consider a system in which the underlying behavioral model is a logit mixture with the utility specification shown in Equation (1). We use the indices n for individuals ($\mathrm{n} = 1, 2, \ldots, \mathrm{N}$), m for menus ($\mathrm{m} = 1, 2, \ldots \mathrm{M_n}$), and j for alternatives ($\mathrm{j} = 1, 2, \ldots, \mathrm{J_{mn}}$).

$$\mathrm{U_{jmn}} = -P_{jmn} + X_{jmn}\beta_n + \frac{1}{\exp\left(\alpha_n\right)}\varepsilon_{\mathrm{jmn}} \tag{1}$$

$$\beta_n \sim \mathrm{N}(\mathrm{B}, \Omega) \tag{2}$$

$\mathrm{U_{jmn}}$ is individual n's utility of alternative j in menu m, P_{jmn} is the price associated with this alternative, $\mathrm{X_{jmn}}$ is a vector of attributes, β_n is a vector of individual-specific parameters following a Normal distribution with population mean B and covariance matrix Ω, $\varepsilon_{\mathrm{jmn}}$ is an error term distributed as Extreme Value I with location zero and scale 1, and α_n is a scale parameter (exponentiation is used to guarantee that the scale is positive for all individuals). The model is specified in the willingness-to-pay (WTP) space; the cost coefficient is fixed to -1 and a scale parameter is estimated. Therefore, all the coefficients represent the WTP values for their corresponding attributes (see Ben-Akiva et al., 2019; Train and Weeks, 2005). Finally, we assume that the alternative-specific constants are included in X_{jmn} as vectors of ones.

In adaptive choice contexts, menus (or choice situations) are generated using the previous choices. Without loss of generality, we assume that:

1. The system is initialized with one or more menus having exogenous attributes $\mathrm{X_{n0}}$. The choices in these menus are denoted as $\mathrm{d_{n0}}$.
2. A known function (Q) is used for generating menu m with alternative attributes $\mathrm{X_{nm}}$ based on the user's previous choices $(\mathrm{d_{n0}}, \ldots, \mathrm{d_{n,m-1}})$ and attributes $(\mathrm{X_{n0}}, \ldots, \mathrm{X_{n,m-1}})$ (where $\mathrm{Q}(\mathrm{X_{nm}} | \mathrm{X_{n0}}, \ldots, \mathrm{X_{n,m-1}}, \mathrm{d_{n0}}, \ldots, \mathrm{d_{n,m-1}})$ represents the probability of menu $\mathrm{X_{nm}}$ conditional on these choices and attributes).

In recommender systems, the Q function is an assortment optimization that generates a personalized menu of alternatives based on the estimated individual-specific coefficients β_n and α_n.

Ideally, the likelihood function used in model estimation should include:

3. The probability of the observed choices conditional on the attributes $\mathrm{P}(\mathrm{d_{nm}} | \mathrm{X_{nm}}, \beta_n, \alpha_n)$ (which depends on the true model parameters).
4. The probability of menu with attributes X_{nm} conditional on the previous choices and attributes $\mathrm{Q}(\mathrm{X_{nm}} | \mathrm{d_{n0}}, \ldots, \mathrm{d_{n,m-1}}, \mathrm{X_{n0}}, \ldots, \mathrm{X_{n,m-1}})$: This is dependent on the true model parameters only through the choices.

The joint likelihood of the choices and attributes (up to menu *m*) is presented in Equation (3).

$$\begin{aligned}&\mathrm{P}(\mathrm{d_{n0}}, \ldots, \mathrm{d_{nm}}, \mathrm{X_{n1}}, \ldots, \mathrm{X_{nm}} | \mathrm{X_{n0}}, \beta_n, \alpha_n) = \mathrm{P}(\mathrm{d_{n0}}\ | \mathrm{X_{n0}}, \beta_n, \alpha_n)\mathrm{Q}(\mathrm{X_{n1}} | \mathrm{d_{n0}}, \mathrm{X_{n0}})\mathrm{P}(\mathrm{d_{n1}} | \mathrm{X_{n1}}, \beta_n, \alpha_n)\ldots\mathrm{Q} \\ &\left(\mathrm{X_{nm}} \middle| \mathrm{d_{n0}}, \ldots, \mathrm{d_{n,m-1}}, \mathrm{X_{n0}}, \ldots, \mathrm{X_{n,m-1}}\right)\mathrm{P}(\mathrm{d_{nm}} | \mathrm{X_{nm}}, \beta_n, \alpha_n)\end{aligned} \tag{3}$$

2.2. Consistent estimation

When a typical model estimation is done, modeling is often limited to the series of choices, and therefore, the likelihood function (of an individual) that is being considered is:

$$\mathrm{P}(\mathrm{d_{n0}}, \ldots, \mathrm{d_{nm}}\ | \mathrm{X_n}, \beta_n, \alpha_n) = \mathrm{P}(\mathrm{d_{n0}}, \mathrm{d_{n1}}, \ldots, \mathrm{d_{nm}} | \mathrm{X_{n0}}, \mathrm{X_{n1}}, \ldots, \mathrm{X_{nm}}, \beta_n, \alpha_n) = \mathrm{P}(\mathrm{d_{n0}} | \mathrm{X_{n0}}, \beta_n, \alpha_n)\mathrm{P}(\mathrm{d_{n1}} | \mathrm{X_{n1}}, \beta_{\mathrm{n}}, \alpha_n)\ldots\mathrm{P}(\mathrm{d_{nm}} | \mathrm{X_{nm}}, \beta_{\mathrm{n}}, \alpha_n) \tag{4}$$

Although this looks different from the likelihood function shown in Equation (3), this is not always problematic. When the full choice history is used, the probability of the recommended alternatives in menu m, conditional on the previous choices, is a constant. For example, in a recommender system with a deterministic function Q, given the previous choices and menus presented to an individual, the recommendation becomes deterministic:

$$\mathrm{Q}(\mathrm{X_{n1}}|\mathrm{d_{n0}},\mathrm{X_{n0}}) = \ldots = \mathrm{Q}\left(\mathrm{X_{nm}}\middle|\mathrm{d_{n0}},\ldots,\mathrm{d_{n,m-1}},\mathrm{X_{n0}},\ldots,\mathrm{X_{n,m-1}}\right) = 1\forall m \tag{5}$$

As a result, these constant terms can be dropped from the likelihood without having any effect on the estimation results:

$$\mathrm{P}(\mathrm{d_{n0}},\ldots,\mathrm{d_{nm}},\mathrm{X_{n1}},\ldots,\mathrm{X_{nm}}|\mathrm{X_{n0}},\beta_n,\alpha_n) = \text{Constant} \times \mathrm{P}(\mathrm{d_{n0}}\ |\mathrm{X_{n0}},\beta_n,\alpha_n)\mathrm{P}(\mathrm{d_{n1}}|\mathrm{X_{n1}},\beta_n,\alpha_n)\ldots\mathrm{P}(\mathrm{d_{nm}}|\mathrm{X_{nm}},\beta_n,\alpha_n) \tag{6}$$

Therefore, the two likelihood functions presented in Equations (3) and (4) are equivalent, and endogeneity is not a concern. This shows that using the entire choice history results in consistent estimation results.

2.3. *Inconsistent estimation*

When the likelihood function does not reflect the data generation process, the obtained estimates might be inconsistent. For example, when the first menu ($\{\mathrm{d_{n0}},\mathrm{X_{n0}}\}$) is excluded from estimation, the correct likelihood function can be derived from Equation (3), by integrating the partial likelihood function over the distribution of $\mathrm{X_{n0}}$ and summing over $\mathrm{d_{n0}}$, as shown in Equation (7):

$$\mathrm{P}(\mathrm{d_{n1}},\ldots\mathrm{d_{nm}},\mathrm{X_{n1}},\ldots,\mathrm{X_{nm}}|\ \beta_n,\alpha_n) = \int_{\mathrm{X_{n0}}}\sum_{\mathrm{d_{n0}}}\mathrm{P}(\mathrm{d_{n0}}|\mathrm{X_{n0}},\beta_n,\alpha_n)\mathrm{Q}(\mathrm{X_{n1}}|\mathrm{d_{n0}},\mathrm{X_{n0}})\mathrm{P}(\mathrm{d_{n1}}|\mathrm{X_{n1}},\beta_n,\alpha_n)\ldots$$

$$\ldots\mathrm{Q}\left(\mathrm{X_{nm}}\middle|\mathrm{d_{n0}},\ldots,\mathrm{d_{n,m-1}},\mathrm{X_{n0}},\ldots,\mathrm{X_{n,m-1}}\right)\mathrm{P}(\mathrm{d_{nm}}|\mathrm{X_{nm}},\beta_n,\alpha_n)\mathrm{f}(\mathrm{X_{n0}})\mathrm{dX_{n0}}$$

$$= \mathrm{Q}\left(\mathrm{X_{n1}},\ldots,X_{nm}\middle|\mathrm{d_{n1}},\ldots,\mathrm{d_{n,m-1}},\beta_n,\alpha_n\right)\mathrm{P}(\mathrm{d_{n1}}|\mathrm{X_{n1}},\beta_n,\alpha_n)\ldots\mathrm{P}(\mathrm{d_{nm}}|\mathrm{X_{nm}},\beta_n,\alpha_n) \tag{7}$$

where $\mathrm{Q}(\mathrm{X_{n1}},\ldots,X_{nm}|\mathrm{d_{n1}},\ldots,\mathrm{d_{n,m-1}},\beta_n,\alpha_n)$ is the marginal probability of observing menus $\mathrm{X_{n1}},X_{n2}\ \ldots,X_{nm}$ conditional on the previous choices:

$$\mathrm{Q}\left(\mathrm{X_{n1}},\ldots,X_{nm}\middle|\mathrm{d_{n1}},\ldots,\mathrm{d_{n,m-1}},\beta_n,\alpha_n\right) = \int_{\mathrm{X_{n0}}}\sum_{\mathrm{d_{n0}}}\mathrm{P}(\mathrm{d_{n0}}|\mathrm{X_{n0}},\beta_n,\alpha_n)\mathrm{Q}(\mathrm{X_{n1}}|\mathrm{d_{n0}},\mathrm{X_{n0}})\ldots\mathrm{Q}\left(\mathrm{X_{nm}}\middle|\mathrm{d_{n0}},\ldots,\mathrm{d_{n,m-1}},\mathrm{X_{n0}},\ldots,\mathrm{X_{n,m-1}}\right)\mathrm{f}(\mathrm{X_{n0}})\mathrm{dX_{n0}} \tag{8}$$

Equation (8) shows that when the first menu is excluded from estimation, $\mathrm{X_{n1}},X_{n2}\ \ldots,X_{nm}$ depend on β_n and α_n, resulting in endogeneity. In this equation, $\mathrm{f}(\mathrm{X_{n0}})$ is determined by the analyst (or the design of the recommender system). The Q terms cannot be treated as constants and dropped from the likelihood as before; they are equal to one only for a specific combination of $\{\mathrm{d_{n0}},\mathrm{X_{n0}}\}$ (which yields the observed values of $\mathrm{X_{n1}},X_{n2}\ \ldots,X_{nm}$), and zero otherwise.

Therefore, when a model is estimated only with a series of choices, similar to Equation (4), the misspecified likelihood function is:

$$\mathrm{P}(\mathrm{d_{n1}},\mathrm{d_{n2}},\ldots,\mathrm{d_{nm}}\ |\mathrm{X_n},\beta_n,\alpha_n) = \mathrm{P}(\mathrm{d_{n1}},\mathrm{d_{n2}},\ldots,\mathrm{d_{nm}}|\mathrm{X_{n1}},\mathrm{X_{n2}},\ldots,\mathrm{X_{nm}},\beta_n,\alpha_n) = \mathrm{P}(\mathrm{d_{n1}}|\mathrm{X_{n1}},\beta_\mathrm{n},\alpha_n)\mathrm{P}(\mathrm{d_{n2}}|\mathrm{X_{n2}},\beta_\mathrm{n},\alpha_n)\ldots\mathrm{P}(\mathrm{d_{nm}}|\mathrm{X_{nm}},\beta_\mathrm{n},\alpha_n) \tag{9}$$

By ignoring the term $\mathrm{Q}(\mathrm{X_{n1}},\ldots,X_{nm}|\mathrm{d_{n1}},\ldots,\mathrm{d_{n,m-1}},\beta_n,\alpha_n)$, this estimation will result in inconsistent estimates.

In conclusion, using the incorrect likelihood function (with only part of the historical data) will result in a dependency of the explanatory variables ($\mathrm{X_{n1}},X_{n2}\ \ldots,X_{nm}$) on the individual-specific parameters (β_n,α_n). The above results can be generalized to conclude that excluding any menu (that is used in generating future recommendations) will result in inconsistent estimates. On the other hand, using all the data might not be feasible, either because of data availability, or because of computational constraints. In order to address the issue of endogeneity, we use a variation of the CF method described in Sections 3.2.

3. Control-function correction for endogeneity

Before presenting our methodology, we provide a brief background of the control function method in Section 3.1, which is used when one or more explanatory variables are correlated with the error term. This has been extensively discussed in the literature. For a more thorough discussion, we refer the reader to Train (2009).

3.1. *Correction for correlation with additive error terms*

The basic version of the CF method addresses correlation between one or more explanatory variables and the error term. For explanatory purposes, we consider first the simple utility depicted in Equation (10):

$$U_{jmn} = \beta_y y_{jmn} + X_{jmn}\beta_X + \varepsilon_{jmn} \tag{10}$$

where y_{jmn} is the endogenous variable that is correlated with the error term (ε_{jmn}), and X_{jmn} is a vector of explanatory variables that are uncorrelated with ε_{jmn}.

The CF method uses exogenous instrumental variables, Z_{jmn}, which are correlated with the endogenous variable y_{jmn}, but not with ε_{jmn} to estimate the residuals in the *first stage* shown in Equation (11).

$$y_{jmn} = E\left(y_{jmn}\middle|Z_{jmn},X_{jmn}\right) + \mu_{jmn}, \tag{11}$$

where the first term, $E(y_{jmn}\Big|Z_{jmn}, X_{jmn})$, is uncorrelated with ε_{jmn} (because it is a function of the instrument Z_{jmn}, and X_{jmn} which are uncorrelated with ε_{jmn}), while the second term, μ_{jmn} retains the endogeneity in y_{jmn}.

The error term (ε_{jmn}) can be assumed to consist of two terms as shown in Equation (12): an extreme value error ($\widetilde{\varepsilon}_{jmn}$), and a normally distributed term that is correlated with y_{jmn}, ζ_{jmn} (Train, 2009).

$$\varepsilon_{jmn} = \widetilde{\varepsilon}_{jmn} + \zeta_{jmn} \tag{12}$$

ζ_{jmn} can be expressed as a function of its mean conditional on μ_{jmn}, denoted by $E(\zeta_{jmn}|\mu_{jmn})$, as shown in Equation (13). This conditional expectation is referred to as the control-function,

$$\zeta_{jmn} = E\left(\zeta_{jmn}|\mu_{jmn}\right) + \nu_{jmn} \tag{13}$$

and is usually assumed to be linear: $E(\zeta_{jmn}\ \Big|\mu_{jmn}) = \lambda\mu_{jmn}$, which follows from assuming that ζ_{jmn} and μ_{jmn} are jointly normal with zero mean and constant covariance matrix for all j (Train, 2009). When the endogenous variable y_{jmn} is continuous, a linear regression is usually used to estimate the residuals, μ_{jmn}. The parameters λ are then estimated along with β_X and β_y by including the control-function in the utility equation, as shown in Equation (14):

$$U_{jmn} = \beta_y y_{jmn} + X_{jmn}\beta_X + \lambda\mu_{jmn} + \nu_{jmn} + \widetilde{\varepsilon}_{jmn} \tag{14}$$

Estimation of a logit mixture model using the utility equation presented in Equation (14) results in consistent estimates of β_X and β_y subject to a scale (and under the linearity assumptions in Equation (13), and the first stage specification in Equation (11)).

3.2. Correction for correlation with individual-specific parameters

The simple case presented above assumes that endogeneity arises as a result of the correlation between the endogenous variable(s) and the unobserved additive error term. However, in the presence of taste heterogeneity, endogeneity can also arise if the individual-specific parameters are correlated with the explanatory variables. In this section, we extend the CF method presented above to the case of random coefficients, and correlations between the attributes and the individual-specific effects. In Section 4, we demonstrate that this methodology can reduce the bias substantially[1] using a Monte Carlo experiment, and result in superior predictions.

3.2.1. Model specification

In order to demonstrate the proposed CF correction, we consider the model shown in Equation (1). We can multiply all the elements in this equation by $\exp(\alpha_n)$, in order to obtain error terms that are independently and identically distributed as Extreme Value EV(0,1), as shown in Equation (15):

$$U_{jmn} = \exp(\alpha_n)\left(-P_{jmn} + X_{jmn}\beta_n\right) + \varepsilon_{jmn}, \tag{15}$$

To obtain consistent estimation results, the distributions of heterogeneity need to be uncorrelated with the covariates (see Wooldridge, 2015). This is because the integration in Equation (16) requires that the distribution $f(\beta_n, \alpha_n|\gamma)$ does not depend on X. Ignoring this dependence results in a correlation between the explanatory variables and the error term, and thus endogeneity bias might arise.

$$P(d|X,\gamma) = \int\limits_{\beta_n,\alpha_n} P(d|X,\beta_n,\alpha_n) f(\beta_n,\alpha_n|\gamma) d\beta_n d\alpha_n, \tag{16}$$

where γ represents hyper-parameters of the joint distribution of β_n and α_n.

Since β_n are normally distributed with means B, we can express Equation (15) as:

$$U_{jmn} = \exp(\alpha_n)\left(-P_{jmn} + X_{jmn}(B + \nu_n)\right) + \varepsilon_{jmn} \tag{17}$$

where ν_n are normally distributed with means 0 and covariance matrix Ω. In this case, endogeneity is expected if the individual-specific parameters, ν_n are correlated with the explanatory variables, X_{jmn}. Such cases are frequently observed in traditional choice contexts (e.g., individuals with a higher time sensitivity tend to relocate closer to their workplace, resulting in shorter travel times). However, our research is motivated by recommender systems,[2] in which individuals with a higher preference towards a certain attribute k (i.e., individuals with a higher value of ν_{kn}) are recommended alternatives having higher values of this attribute.

Our CF correction uses instruments Z_{jmn} that are correlated with X_{jmn}, but not with ν_n. We can carry out the first stage estimation as before, and obtain the residuals μ_{jmn}:

[1] In this paper, we refer to “bias” as the difference between the true values of the model parameters and their corresponding estimates (i.e., finite sample empirical bias).

[2] In recommender systems, the dependence of the distribution of ν_n on X_{jmn} is implied by the recommendation function, Q(X| β_n) (see Equation (8)).

$$X_{jmn} = E\left(X_{jmn} \middle| \mathrm{Z_{jmn}}\right) + \mu_{jmn} \tag{18}$$

In addition to the instrumental variables ($\mathrm{Z_{jmn}}$), this equation includes an intercept and other exogenous variables, if they exist. The individual-specific parameters can be expressed as:

$$\nu_n = E\left(\nu_n \middle| \mu_{jmn}\right) + \nu_n^* \tag{19}$$

Afterwards, we can substitute Equation (19) into the utility equation as shown in Equation (20):

$$U_{jmn} = \exp(\alpha_n)\left(-P_{jmn} + X_{jmn}\left(B + E\left(\nu_n \middle| \mu_{jmn}\right) + \nu_n^*\right)\right) + \varepsilon_{jmn} \tag{20}$$

If we assume ν_n and μ_{jn} to be jointly normal, Equation (19) becomes:

$$\nu_n = \mu_{jn}\lambda + \nu_n^* \tag{21}$$

where μ_{jn} is a vector of residuals over all choice situations (with elements μ_{jmn}; $m = 1, 2, \ldots M_n$), and λ is a vector of coefficients to be estimated. The model can be simplified by restricting the elements of λ to be the same across all menus (assuming that the menus presented to the same individual are similar, which is a reasonable assumption in the long-run operation), which is the same as using the mean of μ_{jn} across all menus ($\overline{\mu}_{jn}$) in Equation (21).

Finally, the modified utility equation can be expressed as:

$$U_{jmn} = \exp(\alpha_n)\left(X_{jmn}\left(B + \nu_n^* + \lambda\overline{\mu}_{jn}\right)\right) + \varepsilon_{jmn} \tag{22}$$

$$= \exp(\alpha_n)\left(X_{jmn}\beta_n^* + X_{jmn}\lambda\overline{\mu}_{jn}\right) + \varepsilon_{jmn}$$

where $\beta_n^* = B + \nu_n^*$, which is not correlated with X_{jmn} (because the instruments Z_{jn} are not correlated with the individual-specific effects). The resulting model is a logit mixture (with random parameters β_n^*), which also includes the control-functions $X_{jmn}\lambda\overline{\mu}_{jn}$ in the utility equations. This extends the standard CF method presented in Section 3.1 by including interactions between the residuals and the explanatory variables. Since X_{jmn} includes a vector of ones (to allow for an intercept), $\lambda\overline{\mu}_{jn}$ will also be included in the utility equations without being interacted with any variables. This is also relevant to recommender systems where we observe correlations between the individual-specific intercepts and one or more variables (for example, in a personalized travel advisor, individuals who prefer transit might be presented with alternatives having a lower transit cost or time).

The model can be generalized as suggested by Wooldridge (2015), Train (2009), and Card (1993) to account for more complicated cases. For example, if there is a nonzero correlation between β_{jn} and $X_{j'mn}$ for $j \neq j'$ (i.e., correlation between the preferences towards one alternative and an attribute of another alternative), then the conditional expectation of β_{jn}^* is modeled as a function of $\overline{\mu}_{j'n}$ (i.e., generally, the residuals for all alternatives enter the utility for each alternative).

3.2.2. Model estimation

The model shown in Equation (22) is a standard logit mixture model, with additional terms in the utility equation multiplied by the residuals. Therefore, the model can be estimated using frequentist estimation procedures, such as simulated maximum-likelihood (MSL) or Bayesian procedures, such as Hierarchical Bayes (HB)/Gibbs sampling (Train, 2009). In this paper, we use the Hierarchical Bayes procedure described in Train (2009), due to its simplicity and computational efficiency.

Bayesian estimation is used to replicate maximum likelihood estimates (and standard errors) at a lower computational cost. The point estimates of the population means and variances are obtained by averaging the posterior draws of these parameters, which pertains to minimizing the quadratic loss function (Koop et al., 2007). Because of the uninformative priors and the large sample size, Bayesian and classical estimates are virtually identical, despite their algorithmic differences (Ben-Akiva et al., 2019; Huber and Train, 2001; Train, 2009), according to the Bernstein-von Mises theorem. Train (2009) shows that the mean of the posterior is equivalent to the maximum likelihood estimator, and the standard deviations of the posterior provide classical standard errors for the estimator. Huber and Train (2001) extend this result to show that the individual-specific parameters obtained from both estimation methods are identical. Therefore, estimation using MSL does not affect the estimation results, and hence the findings of this paper.

The HB estimator is based on a four-step Gibbs sampler with two embedded Metropolis-Hastings algorithms. The Gibbs steps are:

1. Step 1: drawing from the conditional posterior of the population means $\mathrm{B}|\Omega^*, \zeta_n, \lambda$ using a Normal Bayesian update with unknown mean and known variance, where $\zeta_n = [\beta_n^*, \alpha_n] \sim N(\mathrm{B}, \Omega^*)$.[3] The conditional posterior is given by Equation (23):

$$K(\mathrm{B}|\zeta_n \forall n, \Omega^*, \lambda) \propto f(\zeta_n|\mathrm{B}, \Omega^*)k(\mathrm{B}) \tag{23}$$

Where $k(\mathrm{B})$ is a noninformative Normal prior (with a very large variance). A draw from this conditional posterior is obtained by sampling from the distribution $\mathcal{N}\left(\overline{\zeta}^{i-1}, \frac{\Omega^{*i-1}}{N}\right)$, where i is an iteration index, and:

[3] Here we assume that B and Ω^* include also the population mean and variance of the scale parameter.

$$\overline{\zeta}^{i-1} = \frac{1}{N}\sum_n \zeta_n^{i-1} \tag{24}$$

2. Step 2: drawing from the conditional posterior of the covariance matrix $\Omega^*|\mathrm{B},\zeta_n$ using a Normal Bayesian update with known mean and unknown variance. The conditional posterior on Ω^* is given by:

$$K(\Omega^*|\mathrm{B},\zeta_n\forall n,\lambda) \propto f(\zeta_n|\mathrm{B},\Omega^*)k(\Omega^*) \tag{25}$$

Where $k(\Omega^*)$ is a Hierarchical Inverted Wishart prior (Huang and Wand, 2013). A draw from this conditional posterior is obtained by sampling from the distribution $IW(\nu + N + K - 1, \overline{V} + 2\nu D)$, where:

$$\overline{V} = \frac{1}{N}\sum_n \left(\zeta_n^{i-1} - B^i\right)\left(\zeta_n^{i-1} - B^i\right)' \tag{26}$$

and D is a diagonal matrix with diagonal elements element λ_{i} which follow the inverse Gamma distribution $\lambda_{\mathrm{k}} \sim \mathrm{IG}\left(\frac{1}{2},\frac{1}{\mathrm{a}_{\mathrm{k}}^2}\right)$.

The value of ν is recommended to be 2 (Huang and Wand, 2013) and a_{i} are parameters specifying how informative the distribution is, where larger values lead to arbitrarily weakly informative priors on the corresponding standard deviation term.

3. Step 3: drawing from the conditional posterior of the individual-specific parameters $\zeta_n|\mathrm{B},\Omega^*,\lambda$ using the Metropolis-Hastings (MH) algorithm. For simplicity, we use the random-walk MH algorithm defined in Train (2009).

A draw from the conditional posterior, $\widetilde{\zeta}_n^i$, is obtained as $\widetilde{\zeta}_n^i = \zeta_{\mathrm{n}}^{i-1} + \rho L\eta$, where L is the Choleski factor of Ω^{*i}, η are draws from the standard normal density, and ρ is a scalar specified by the researcher, and adjusted within the iterative process to achieve a desirable acceptance rate.

The draw is accepted if:

$$\frac{L\left(d_n\middle|\widetilde{\zeta}_n^i,\lambda^{i-1}\right)\varphi\left(\widetilde{\zeta}_n^i\middle|B^i,\Omega^{*i}\right)}{L\left(d_n\middle|\zeta_{\mathrm{n}}^{i-1},\lambda^{i-1}\right)\varphi\left(\zeta_{\mathrm{n}}^{i-1}\middle|B^i,\Omega^{*i}\right)} \geq u \tag{27}$$

Where u is a draw from the standard uniform distribution. If the new draw is rejected, the previous draw, ζ_{n}^{i-1}, is maintained.

4. Step 4: drawing from the conditional posterior of the coefficients that do not vary among individuals, such as λ, using another Metropolis-Hastings (MH) algorithm.

A draw from the conditional posterior, $\widetilde{\lambda}^i$, is obtained as $\widetilde{\lambda}^i = \lambda^{i-1} + \rho_\lambda\eta_\lambda$, where η_λ is a draw from the standard normal density, and ρ_λ is a scalar specified by the researcher, and adjusted within the iterative process to achieve a desirable acceptance rate.

The draw is accepted if:

$$\frac{L\left(d_n\middle|\zeta_{\mathrm{n}}^i,\widetilde{\lambda}^i\right)}{L\left(d_n\middle|\zeta_{\mathrm{n}}^i,\lambda^{i-1}\right)} \geq u \tag{28}$$

Where u is a draw from the standard uniform distribution. If the draw is rejected, the previous draw, λ^{i-1}, is maintained.

In some cases, the number of parameters in λ might be large, due to the presence of multiple endogenous variables. For example, in the example we show in Section 4.1, there are 60 additional coefficients to be estimated. Sampling from all of these 60 coefficients at once in Step 4 might slow down the convergence of the Gibbs sampler. There are two ways in which this can be addressed (however, it is necessary to estimate a full covariance matrix using either method):

1. Carrying out Step 4 in batches, by dividing the vector of λ's into multiple vectors, and sampling from each vector separately, or
2. Modifying Step 1 to include a multivariate linear regression instead of a Normal Bayesian update. This way, we can rewrite the model as:

$$U_{jmn} = \exp(\alpha_n)\left(X_{jmn}\delta_n\right) + \varepsilon_{jmn} \tag{29}$$

where:

$$\delta_n = \beta_n^* + \lambda\overline{\mu}_{jn} = B + \lambda\overline{\mu}_{jn} + \nu_n^* \tag{30}$$

We need to draw from the conditional posterior of $\Lambda = [B,\lambda]$. Therefore, we construct a matrix of explanatory variables as:

$$X_\mu = \begin{bmatrix} 1 & \cdots & \overline{\mu}_{J1} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & \overline{\mu}_{Jn} \end{bmatrix} \tag{31}$$

And the individual-specific parameters are specified as a matrix of independent variables:

$$\varDelta = \begin{bmatrix} \delta_{11} & \cdots & \delta_{1K} \\ \vdots & \ddots & \vdots \\ \delta_{N1} & \cdots & \delta_{NK} \end{bmatrix} \tag{32}$$

Then we estimate the vector of coefficients, $\widehat{\varLambda}$, as $(X_\mu^{'}\mathrm{X}_\mu)^{-1}X_\mu^{'}\Delta$. We obtain a draw from the conditional posterior on $vec(\varLambda^{\mathrm{i}})$ as $N(vec(\widehat{\varLambda}),\varOmega^{*\mathrm{i}-1}\bigotimes(X_\mu^{'}\mathrm{X}_\mu)^{-1})$. We carry out Steps 2 and 3 as before, but using the fitted values of δ_{n}^{i} to calculate $\overline{V}$ in Equation (26) instead of B^i (and we adjust the degrees of freedom of the posterior distribution to reflect the dimensionality of the multivariate regression model). We also use these fitted values instead of B^i in Equation (27).

3.2.3. Model application

This method is able to recover the true population means for the random coefficients (which are the same as the means of β_n^*). However, the variances of β_n are not directly recovered because we estimate the variances of β_n^* (by conditioning on the control − function). Since β_n is expressed as $\beta_n^* + \lambda\overline{\mu}_{jn}$, the variability in β_n is partially explained by $\lambda\overline{\mu}_{jn}$ (which is included in the systematic part of the utility equation). Therefore, we expect the estimated variances to be smaller than the true variances of β_n (However, the variances of β_n can be imputed by adding $var(\lambda\overline{\mu}_{jn})$ to the estimated variances of β_n^*).

This implies that we need to use the control-function residuals in forecasting and model application as proposed by *(*Guevara and Ben-Akiva, 2012*)*. Model application is performed using Equation (33):

$$P(j_{mn}=1|\widehat{B},\widehat{\varOmega}^*,\widehat{\lambda}) = \int P\left(j_{mn}=1|\beta_n^*,\alpha_n,\widehat{\lambda}\right)f\left(\beta_n^*,\alpha_n|\widehat{B},\widehat{\varOmega}^*\right)d\beta_n^* d\alpha_n \tag{33}$$

where $\widehat{B}$ and $\widehat{\varOmega}^*$ are the estimated mean and covariance matrix of β_n^*, and:

$$P\left(j_{mn}=1|\beta_n^*,\alpha_n,\widehat{\lambda}\right) = \frac{\exp\left(V_{jmn}^*\right)}{\sum_{i=1}^{J_{mn}}\exp\left(V_{imn}^*\right)} \tag{34}$$

and:

$$V_{jmn}^* = \exp(\alpha_n)\left(-P_{jmn} + X_{jmn}\beta_n^* + X_{jmn}\,\widehat{\lambda}\overline{\mu}_{jn}\right) \tag{35}$$

3.2.4. Model consistency

Model assumptions: According to Wooldridge (2015), the consistency of the CF estimator depends on two assumptions: the first stage being correctly specified (Equation (18)), and the linearity assumptions (Equation (21)), which follow from the joint normality assumption of the residuals and the individual-specific effects. Previous research using a similar method in the context of a dynamic model with non-linear panel data found that the control-function correction substantially reduces bias, but does not completely eliminate it (Orme, 2001).

The proposed method is mechanically similar to the Wooldridge method (Wooldridge, 2005), which is used to address the initial conditions problem in nonlinear dynamic panel data mixture models, using an auxiliary distribution of the unobserved individual effects which is conditioned on the initial values and exogenous variables. However, in our case, the auxiliary distribution uses the estimated residuals instead of the initial values. According to Akay (2012), the Wooldridge method works very well for long panels (more than 5–8 periods), but still results in significant bias for shorter panels.

Model Instruments: In order to apply our CF correction to recommender systems, we need relevant instruments that are correlated with the endogenous explanatory variables (i.e., the attributes of the recommended alternatives), but not with the individual-specific parameters. In choice-based recommender systems, we propose to use *non-personalized recommendations* as instruments, representing what we would recommend to an *average individual* (from the same universal sets), without observing their previous choices. These instruments can be easily obtained from the recommendation engine, and satisfy both conditions:

- Exogeneity: since the estimated user-specific parameters are not used in generating these instruments. Unlike the endogenous attributes, these instruments are not generated by the function $\mathrm{Q}(X_{nm}|\mathrm{d}_{\mathrm{n}1},\ldots,\mathrm{d}_{\mathrm{n,m}-1},\beta_n,\alpha_n,I_{nm})$, but rather by a similar function $\mathrm{Q}(Z_{nm}|\widehat{B},\widehat{\varOmega},I_{nm})$, where I_{nm} is the initial set or inventory of items to recommend from. There are different tests for instrument exogeneity proposed in the literature, including the overidentification-based refutability and modified refutability tests, and the Hausman test (Guevara, 2018; Hausman, 1978). These tests were originally developed for the correlation with the error term as discussed in Section 3.1, but can be extended in future work to account for the correlation with individual-specific effects.

- Relevance: non-personalized recommendations are correlated with the endogenous attributes (because they are obtained from the same initial set or inventory, I_{nm}). However, in some cases (e.g., if I_{nm} is a large set), weak instruments can result in inconsistent estimation and, consequently, invalid statistical inference. To test for this issue, Frazier et al. (2020) proposed a novel test that can consistently detect weak identification in commonly applied discrete choice models.

4. Monte Carlo experiments

In this section, we present two examples of applying the above methodology to simulated and real datasets. In the first example, we use simulated data to show how our control-function correction addresses endogeneity bias. We carry out 30 repetitions, holding the sample size and the length of the choice history constant, in order to show that the estimated coefficients are generally centered around the true values of the parameters. In the second example, we vary the sample size and the length of the choice history to analyze how sensitive are our estimates to these two factors.

4.1. *Example 1: personalized recommendations using Monte Carlo data*

In the first example, we simulate a personalized recommender system following the methodology of (Danaf et al., 2019; Song et al., 2017, 2018), mimicking the choice of Mobility-as-a-Service (MaaS) plans. This system recommends the "best" alternatives to a user (e.g., alternatives that are most likely to be chosen), based on the estimated individual-specific parameters. In a similar application, Danaf et al. (2020) showed that the estimates are only consistent if the entire choice history of each individual is included in estimation. In the following section, we assume that some observations are excluded from estimation, and then we apply the CF method presented in Section 3.2 to correct for endogeneity.

4.1.1. *Data set description*

The Monte Carlo data assumes that 10,000 individuals are presented with 12 successive menus. A sufficiently large sample size is used because we are interested in analyzing the consistency of the estimates (lower samples result in larger standard errors, which makes it difficult to distinguish bias from statistical discrepancies and simulation errors). While this sample is large compared to traditional applications, it is typical in web or app-based contexts such as recommender systems (where the number of users is large). In addition, endogeneity bias is not a concern when dealing with small samples, as we can usually include the entire choice history in estimation.

Each menu/choice situation includes three alternatives (different MaaS plans) with varying attributes and an opt-out alternative. Each plan has a different monthly price and three attributes: access to transit, access to bike sharing, and the number of on-demand trips (e.g., taxi, Uber, Lyft etc.) per month. Table 1 shows the distributions of the attributes in the data. The dependent variable is the choice between the three different plans or opting-out (indicating that the individual does not purchase any of the MaaS plans, or chooses an outside alternative).

The utility equations (normalized to the opt-out alternative) are presented in Equation (36):

$$\mathrm{U_{jmn}} \equiv -1 \times \mathrm{P_{jmn}} + \beta_{\mathrm{T,n}}\mathrm{T_{jmn}} + \beta_{\mathrm{B,n}}\mathrm{B_{jmn}} + \beta_{\mathrm{D,n}}\mathrm{D_{jmn}} + \beta_{\mathrm{q,n}} + \frac{1}{\exp\,(\alpha_{\mathrm{n}})}\varepsilon_{\mathrm{jmn}}$$

$$\mathrm{U_{opt-out,mn}} \equiv 0 + \frac{1}{\exp\,(\alpha_{\mathrm{n}})}\varepsilon_{\mathrm{opt-out,mn}} \tag{36}$$

where:

- n is an index for users ($\mathrm{n} = 1, 2, \dots, \mathrm{N}$), m is an index for menus ($\mathrm{m} = 1, 2, \dots, \mathrm{M_n}$), and j is an index for alternatives in the menu ($\mathrm{j} = 1, 2, \dots, \mathrm{J_{mn}}$).
- $\mathrm{U_{jmn}}$ represents the utility of alternative j in menu m faced by individual n, and $\mathrm{U_{opt-out,mn}}$ is the opt-out utility.
- $\mathrm{P_{jmn}}$ is the monthly price (in \$100's) of alternative j in menu m faced by individual n, with its coefficient normalized to −1.
- $\mathrm{T_{jmn}}$, $\mathrm{B_{jmn}}$ and $\mathrm{D_{jmn}}$ represent access to transit, access to bike sharing, and the number of on-demand trips per month of alternative j with coefficients $\beta_{\mathrm{T,n}}$, $\beta_{\mathrm{B,n}}$, and $\beta_{\mathrm{D,n}}$ respectively.
- $\beta_{\mathrm{q,n}}$ is a constant term for choosing any plan (rather than opting out).
- α_{n} is a scale parameter (exponentiation used to guarantee that the scale is positive).

Table 1
Monte Carlo attributes and levels.

Attribute	Symbol	Levels
Monthly Price	P	\$0 to \$480[a]
Transit	T	Available (1) or unavailable (0)
Bike Sharing	B	Available (1) or unavailable (0)
On-Demand Trips	D	2 to 12 trips/month

[a] Price is positively correlated with all other attributes.

- $\varepsilon_{\mathrm{jmn}}$ is an error component following the extreme value distribution $(\mathrm{EV}(0,1))$.

The model uses the money-metric utility specification as in Ben-Akiva et al. (2019), whereby the price coefficient is fixed to -1 and a scale parameter is estimated. This specification is advantageous because in it all other coefficients represent the willingness-to-pay for the corresponding attributes (Train and Weeks, 2005), and because we can directly distinguish the impact of endogeneity on the ratio of the model coefficients (Guevara and Ben-Akiva, 2012). Since the price coefficient is fixed, the scale parameter α_n can be estimated. We can multiply all the elements of Equation (36) by $\exp(\alpha_{\mathrm{n}})$ in order to obtain Equation (37), where the error term $\varepsilon_{\mathrm{jmn}}$ is distributed as EV(0,1).

$$\begin{aligned} &\mathrm{U_{jmn}} \equiv \exp(\alpha_n) \times \left(-1 \times \mathrm{P_{jmn}} + \beta_{\mathrm{T,n}}\mathrm{T_{jmn}} + \beta_{\mathrm{B,n}}\mathrm{B_{jmn}} + \beta_{\mathrm{D,n}}\mathrm{D_{jmn}} + \beta_{\mathrm{q,n}}\right) + \varepsilon_{\mathrm{jmn}} \\ &\mathrm{U_{opt-out,mn}} \equiv 0 + \varepsilon_{\mathrm{opt-out,mn}} \end{aligned} \tag{37}$$

All coefficients are normally distributed in the sample. The true values of the population means, and inter-consumer covariance matrix, are shown in Table 2. The true values of the individual specific coefficients are generated only once from their corresponding distributions, and then used in generating the choices in each menu. However, the attributes of the alternatives are different across different individuals and different menus.

The choices are simulated by calculating the systematic utilities (using the true individual-specific coefficients and the attributes) and adding EV(0,1) error terms to these systematic utilities to obtain the total utilities. The alternative with the highest total utility is chosen.

4.1.2. *Personalized recommendations*

In this section, we use the methodology proposed by Danaf et al. (2019, 2020) and Song et al. (2017, 2018) in order to generate personalized recommendations. We assume that each individual is presented with 12 menus, however, we only have access to the last four menus. In these menus, the “best” three alternatives are recommended from the universal set. We use menus $9, 10,$ and 11 for estimation, and the last menu (12) for validation and model application (discussed in Section 4.1.5)

The recommended alternatives are determined by the individual-specific parameters estimated using the previous menus (which are assumed to be unavailable). These parameters can be obtained using the Hierarchical Bayes (HB)) estimator in Train (2009). We first calculate the systematic utility of each alternative in the universal set as per Equation (38), and then recommend the three alternatives with the highest systematic utilities.

$$\begin{aligned} &\widehat{\mathrm{V}}_{\mathrm{jmn}} \equiv \exp(\widehat{\alpha}_n)\left(-\mathrm{P_{jmn}} + \widehat{\beta}_{\mathrm{T,n}}\mathrm{T_{jmn}} + \widehat{\beta}_{B,n}\mathrm{B_{jmn}} + \widehat{\beta}_{D,n}\mathrm{D_{jmn}} + \widehat{\beta}_{\mathrm{q,n}}\right) \quad ; \\ &\mathrm{j} = 1, 2, \ldots, 10, m = 9, 10, 11 \end{aligned} \tag{38}$$

In each menu, the choice among the three MaaS plans and opting-out is simulated using the true individual-specific parameters (and not their estimates).

4.1.3. *Control-function correction*

As explained in Section 3.2, we use the attributes of non-personalized recommendations as instruments. In order to obtain better results, we treat the *ranked* recommendations for each menu as different alternatives and use the ranked instruments as well. For example, P_{1mn} represents the price of the first (best) recommendation, and P_{3mn} represents the price for the third (worst) recommendation. We denote the instruments by $P_{1mn}, P_{2mn}, \ldots, D_{2mn}, D_{3mn}$. This indicates that D_{1mn} represents the on-demand trips corresponding to the best recommendation in the non-personalized menu, and D_{3mn} represents the on-demand trips corresponding to the worst recommendation in this menu.

In the first stage, we regress each of the endogenous variables $(P_{1mn}, P_{2mn}, \ldots, D_{2mn}, D_{3mn})$ on all the instruments $(P_{1mn}, P_{2mn}, \ldots, D_{2mn}, D_{3mn})$. For the binary variables (transit and bike-sharing access), we use probit regressions and calculate the *generalized residuals* (Wooldridge, 2015) using the Inverse Mills Ratios obtained from the probit estimates.

The first stage regressions result in 12 sets of residuals (3 utility equations and 4 endogenous variables) denoted by $(\widehat{\mu}_{P1mn}, \widehat{\mu}_{P2mn}, \ldots, \widehat{\mu}_{D3mn})$, which are averaged over all menus of each individual to obtain $\overline{\mu}_{P1n}, \overline{\mu}_{P2n}, \ldots, \overline{\mu}_{D3n}$. These averaged residuals can be used to model the conditional expectation of each individual-specific parameter, as shown in Equation (39), showing an example for the conditional

Table 2
True values of the parameters.

Parameter	True mean	Covariances				
		$\boldsymbol{\beta}_{T,\mathbf{n}}$	$\boldsymbol{\beta}_{B,\mathbf{n}}$	$\boldsymbol{\beta}_{D,\mathbf{n}}$	$\boldsymbol{\beta}_{q,\mathbf{n}}$	$\boldsymbol{\alpha}_n$
$\boldsymbol{\beta}_{T,\mathbf{n}}$	1	2	0.7	−0.5	0	0
$\boldsymbol{\beta}_{B,\mathbf{n}}$	0.5	0.7	1.2	−0.3	0	0
$\boldsymbol{\beta}_{D,\mathbf{n}}$	1.5	−0.5	−0.3	1.5	0	0
$\boldsymbol{\beta}_{q,\mathbf{n}}$	−0.5	0	0	0	1	0
$\boldsymbol{\alpha}_n$	−0.75	0	0	0	0	0.25

expectation of the transit parameter:

$$\begin{aligned}\beta_{T,n} &= B_T + \nu_{T,n}\\ &= B_T + E\left(\nu_{T,n}\middle|\overline{\mu}_{P1n}, \overline{\mu}_{P2n}, \ldots, \overline{\mu}_{D3n}\right) + \nu^*_{T,n}\\ &= B_T + \lambda_{T1}\overline{\mu}_{P1n} + \lambda_{T2}\overline{\mu}_{P2n} + \ldots + \lambda_{T12}\overline{\mu}_{D3n} + \nu^*_{T,n}\\ &= \beta^*_{T,n} + \lambda_{T1}\overline{\mu}_{P1n} + \lambda_{T2}\overline{\mu}_{P2n} + \ldots + \lambda_{T12}\overline{\mu}_{D3n}\end{aligned} \tag{39}$$

where $\beta^*_{T,n} = B_T + \nu^*_{T,n}$ is distributed across individuals with mean and variance to be estimated, and $\lambda_{T1}, \lambda_{T2}, \ldots, \lambda_{T12}$ are parameters to be estimated as well. We also estimate a full covariance matrix for the distributed parameters $\{\beta^*_{T,n}, \beta^*_{B,n}, \beta^*_{D,n}, \beta^*_{q,n}\}$. The modified utility equations are given by Equation (40):

$$\begin{aligned}&V_{\mathrm{jmn}} \equiv \exp\left(\alpha^*_{\mathrm{n}} + \lambda_{\alpha 1}\overline{\mu}_{P1n} + \lambda_{\alpha 2}\overline{\mu}_{P2n} + \ldots + \lambda_{\alpha 12}\overline{\mu}_{D3n}\right)\times\\ &\Big(-\mathrm{P}_{\mathrm{jmn}} + \beta^*_{T,n}\mathrm{T}_{\mathrm{jmn}} + \beta^*_{B,n}\mathrm{B}_{\mathrm{jmn}} + \beta^*_{D,n}\mathrm{D}_{\mathrm{jmn}} + \beta^*_{q,n} +\\ &T_{jmn} \times (\lambda_{T1}\overline{\mu}_{P1n} + \lambda_{T2}\overline{\mu}_{P2n} + \ldots + \lambda_{T12}\overline{\mu}_{D3n}) +\\ &B_{jmn} \times (\lambda_{B1}\overline{\mu}_{P1n} + \lambda_{B2}\overline{\mu}_{P2n} + \ldots + \lambda_{B12}\overline{\mu}_{D3n}) +\\ &D_{jmn} \times (\lambda_{D1}\overline{\mu}_{P1n} + \lambda_{D2}\overline{\mu}_{P2n} + \ldots + \lambda_{D12}\overline{\mu}_{D3n}) +\\ &\lambda_{Q1}\overline{\mu}_{P1n} + \lambda_{Q2}\overline{\mu}_{P2n} + \ldots + \lambda_{Q12}\overline{\mu}_{D3n}\Big) \quad ; \; j = 9, 10, 11 \qquad (40)\\ &V_{\mathrm{opt-out,mn}} \equiv 0\end{aligned}$$

Note that we also express the scale and the constant as functions of the residuals, because these are also individual-specific parameters that can be susceptible to endogeneity.

4.1.4. Results

The model was estimated using the Gibbs sampler described in Section 3.2.2. A random walk MH was used in the third and fourth layers of the Gibbs sampler. We used two million Gibbs iterations which were more than enough to guarantee convergence (according to the Gelman and Rubin's convergence diagnostic, and the Heidelberger and Welch's convergence diagnostic, which are available in the R package 'coda', by Plummer et al., 2020). The first million draws were discarded, and each 10,000th draw from the second million was saved, to arrive at a final sample of 1000 posterior draws for each estimated parameter.

The results presented in Table 3 indicate that the CF correction can recover the population means of the transit, bike-sharing, and on-demand parameters, in addition to the intercept, as expected. On the other hand, if we do not apply any correction, the parameters of transit, bike-sharing, and on-demand will be biased upwards, and statistically different from the true values at the 95% level of confidence. We also note that the magnitude of bias (with no correction) decreases as more menus are included. This is consistent with the empirical results found by Danaf et al. (2020), Fowkes (2007), Abernethy et al. (2007), and Akay (2012).

On the other hand, the estimated variances are smaller than the true values as expected (see Section 3.2). This is because the estimated variances are those of β^*_n, and not β_n. Since β_n is expressed as $\beta^*_n + \lambda\overline{\mu}_{jn}$, the variability in β_n is partially explained by $\lambda\overline{\mu}_{jn}$,

Table 3
Estimation results: posterior means (and standard deviations in parentheses) with and without the CF correction.

Population Means					
	True Means	No Correction (1 menu)	CF (1 menu)	No Correction (3 menus)	CF (3 menus)
Transit	1.0	2.098 (0.085)	1.105 (0.058)	1.131 (0.034)	0.994 (0.024)
Bike Sharing	0.5	1.320 (0.093)	0.509 (0.057)	0.604 (0.031)	0.480 (0.022)
On-Demand	1.5	2.646 (0.125)	1.489 (0.083)	1.694 (0.045)	1.510 (0.033)
Constant	−0.5	−2.671 (0.263)	−0.551 (0.107)	−0.342 (0.054)	−0.515 (0.038)
Scale	0.75	0.033 (0.062)	0.981 (0.093)	0.473 (0.021)	0.724 (0.023)
Population Variances					
	True Variances	No Correction (1 menu)	CF (1 menu)	No Correction (3 menus)	CF (3 menus)
Transit	2.0	3.444 (0.361)	1.08 (0.203)	1.781 (0.097)	0.763 (0.050)
Bike Sharing	1.2	2.329 (0.264)	1.132 (0.162)	1.324 (0.080)	0.719 (0.042)
On-Demand	1.5	3.887 (0.682)	1.174 (0.211)	1.548 (0.135)	0.653 (0.067)
Constant	1.0	32.914 (3.058)	1.067 (0.264)	2.852 (0.236)	0.957 (0.094)
Scale	0.25	1.066 (0.117)	1.714 (0.335)	0.764 (0.069)	0.537 (0.064)

which results in a smaller unobserved variance. Therefore, the variances shown in Table 3 cannot be directly compared to the true values shown in Table 2. Even though these parameters are not estimated consistently, this is not problematic because the estimated residuals can be used in forecasting, as discussed in Section 3.2.3. This is demonstrated in Section 4.1.5.

In order to show that the results presented in Table 3 are not accidental, we present the distribution of the estimated means obtained from 30 replications from the same data generation process in Figs. 1 and 2. These results indicate that the distributions obtained with the CF correction are usually centered around the true values of the estimates, while those obtained without any correction are not.

Finally, in order to estimate the population covariance matrix, we need to account for the variances and covariances implied by Equation (39). The variances presented in Table 3 correspond only to β_n^*. The population variances can thus be approximated by adding the estimated variances to the variance of $\lambda\overline{\mu}_{jn}$ (because they are uncorrelated by construction). The results (with 3 menus) are shown in Table 4.

4.1.5. Model application

In this section, we apply the estimated model to out-of-sample data using the procedure explained in Section 3.2. We use the estimation results with three choice situations per individual (9, 10, 11) in order to predict the choice in the last menu/choice situation. Our analysis is based on the log-likelihood and the average predicted probability of the chosen alternative. The results are presented in Table 5. In calculating the log-likelihood, we used the posterior means of the estimated parameters (as mentioned in Section 3.2, Bayesian estimation is used to replicate maximum likelihood estimates by using uninformative priors). In estimating the conditional choice probability and hit-rate, we used all the draws from the posterior distributions of the individual-specific parameters.

The results indicate that the predictions obtained using the control-function correction are substantially better than those obtained without any correction. In addition, these predictions are even better than those obtained using the true values of the population means and the covariance matrix. This is because the residuals explain part of the random taste variation, which exploits the information in the attributes. These results are similar to the findings of Danaf et al. (2019b), Bhat (2000), and Horsky et al. (2006), who found that models with systematic taste heterogeneity (i.e., preference heterogeneity that is explained by covariates) outperform models with random taste heterogeneity in terms of predictions.

Finally, we use our estimates of the individual-specific parameters from the three menus to predict the estimated hit-rate, defined as the percentage of cases where individuals chose the alternative with the highest predicted probability. In addition, we compare our predictions to those obtained from the individual-specific parameters obtained by conditioning on the true values of the mean and

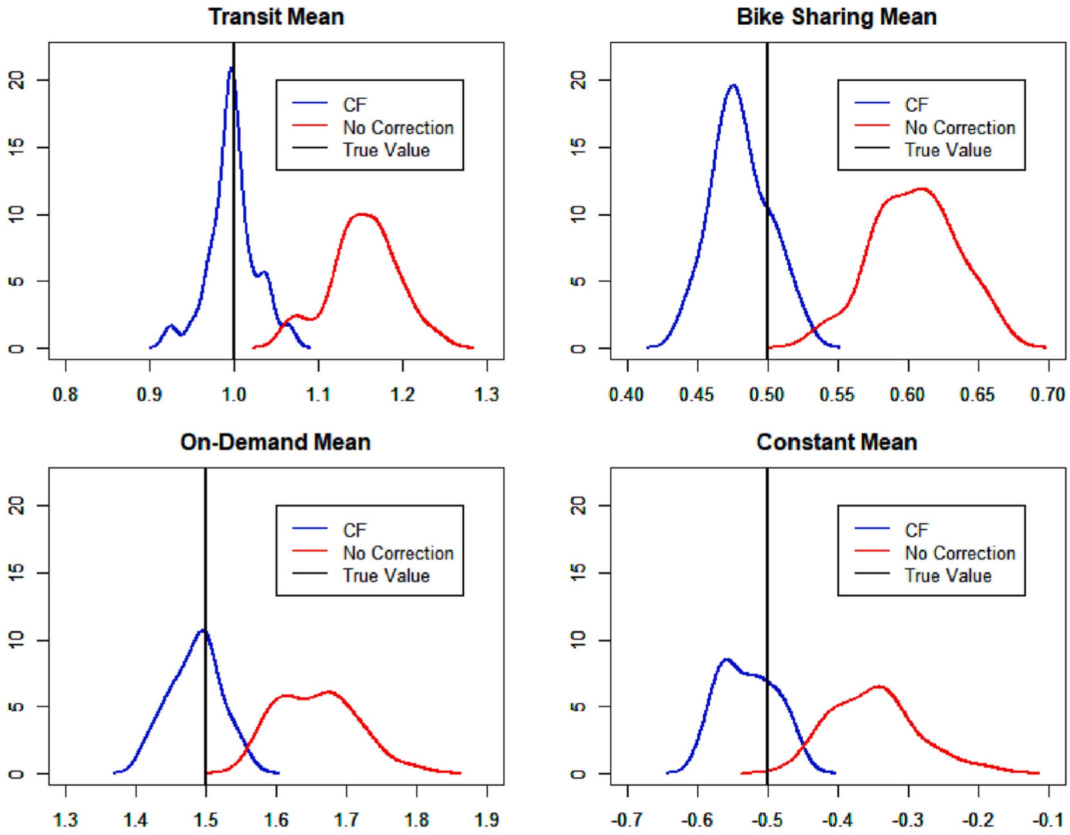

Fig. 1. Distribution of the estimates of the population means with and without the CF correction (obtained with 3 menus per individual).

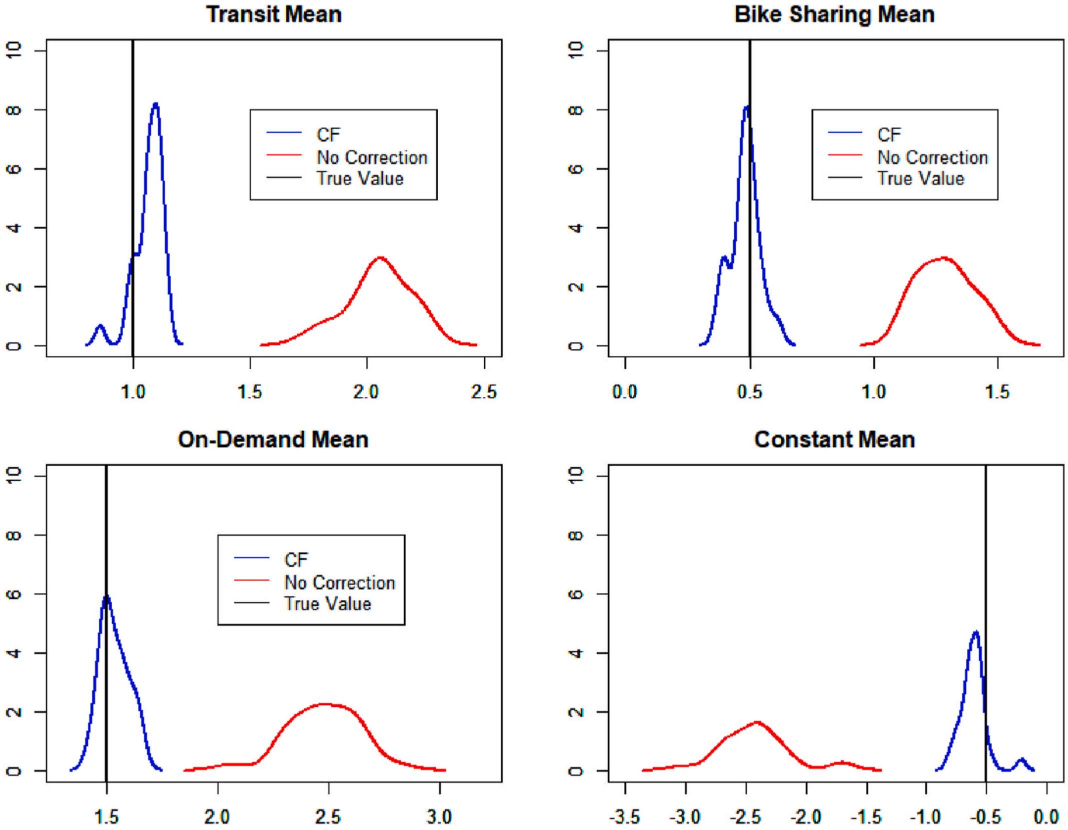

Fig. 2. Distribution of the estimates of the population means with and without the CF correction (obtained with 1 menu per individual).

Table 4
Estimating the covariance matrix (posterior means).

	True Covariance Matrix						Estimated Covariance Matrix				
	$\beta_{T,n}$	$\beta_{B,n}$	$\beta_{D,n}$	$\beta_{q,n}$	α_n		$\beta_{T,n}$	$\beta_{B,n}$	$\beta_{D,n}$	$\beta_{q,n}$	α_n
$\beta_{T,n}$	2	0.7	−0.5 0	0	0	$\beta_{T,n}$	1.902	0.632	−0.379	−0.079	0.003
$\beta_{B,n}$	0.7	1.2	−0.3	0	0	$\beta_{B,n}$	0.632	1.316	−0.273	−0.123	0.001
$\beta_{D,n}$	−0.5	−0.3	1.5	0	0	$\beta_{D,n}$	−0.379	−0.273	1.506	0.036	−0.096
$\beta_{q,n}$	0	0	0	1	0	$\beta_{q,n}$	−0.079	−0.123	0.036	1.033	−0.092
α_n	0	0	0	0	0.25	α_n	0.003	0.001	−0.096	−0.092	0.548

Table 5
Predictions on the third menu using models estimated on the first two menus.

	Log-Likelihood	Mean Prob.	Mean Conditional Prob.	Hit-Rate
Control Function	−7 128.0	0.575	0.686	0.741
No Correction	−11065.5	0.400	0.561	0.617
True Values	−10966.2	0.404	0.741*	0.810*

* The true individual-specific parameters are used in estimating these two values.

covariance matrix and the observed choices. The results in Table 5 also indicate that the CF predictions perform better than the uncorrected model.

4.2. Example 2: personalized advertisement using a real dataset

In the second example, we present a case study with a real dataset obtained from the mlogit R package (Croissant, 2014). However, we modify the data to introduce endogeneity in the form of personalized advertisement. The original dataset (Jain et al., 1994) contains annual scanner data on the purchasing choice between Heinz and Hunts tomato catsup for a panel of 300 households over approximately two years. Heinz is the major brand with three different sizes, Heinz 40, Heinz 32, and Heinz 28, with market shares of 0.065, 0.521, and 0.304, respectively. The other brand is Hunt's 32 with a share of 0.110. In addition to the price of each alternative, the dataset includes two variables: (1) whether there is a special display for brand *j* (labelled as $disp_j$) and whether brand *j* had a featured advertisement in the newspaper ($feat_j$).

4.2.1. Model estimation

As in the original paper, we consider the three sizes of Heinz and as three different brands. We use the random coefficients choice model shown in Equation (41) to estimate the initial model. The estimation results are shown in Table 6.

$$U_{jnm} = -p_{jnm} + ASC_{jn} + \beta_{disp,n} * disp_{jnm} + \beta_{feat} * feat_{jnm} + \frac{1}{\exp\ (\alpha)}\varepsilon_{\mathrm{jmn}} \tag{41}$$

where:

- U_{jnm} is household n's random utility of alternative j in choice situation m.
- p_{jnm} is the price of alternative j in choice situation m seen by household n's. As before, the price coefficient is fixed to -1 (money-metric utility) and a scale parameter (α) is estimated.
- $disp_{jnm}$ and $feat_{jnm}$ are binary variables indicating whether alternative j had a special display or newspaper feature.
- ASC_{jn} represent household-specific alternative-specific constants (ASCs). The ASC of Hunts catsup is fixed to 0.
- $\beta_{disp,n}$ is a random household-specific coefficient for display. A full covariance matrix is used for the parameters ASC_{jn} and $\beta_{disp,n}$.
- β_{feat} and α are the newspaper feature coefficient and the scale parameter respectively. These were not treated as random parameters, because heterogeneity was found to be insignificant.

In the initial model estimation, the posterior means of the population variances of the "Featured" coefficient and the scale were very close to zero, but their posterior standard deviations were large. In a frequentist context, these variances would be deemed insignificant. Therefore, we have dropped them from the model.

4.2.2. Simulation with endogeneity

We modify the dataset to introduce endogeneity by making the "display" attribute a function of the household's previous choices, mimicking personalized advertisement. The underlying assumption is that the displayed brands for a particular household are those that are most likely of being purchased by this household.

We use the original dataset to generate simulated data, where we assume that the coefficients estimated in Table 6 are the true values. Using these preference distributions, we generate a sample of 5000 households. For these households, we generate two synthetic datasets: (1) an exogenous dataset (6 choices per household) used to estimate the household-specific parameters, and (2) an endogenous dataset (5 choices per household) in which the display variable is determined by the household's estimated preferences. The display attribute is set to 1 for the two alternatives that are most likely to be chosen by a particular household, based on the household's estimated parameters from the initial dataset. For each household *n*, we calculate the systematic utility of each brand (without the display variable) as shown in Equation (42). Afterwards, the two brands with the highest systematic utilities are assigned a display value of 1, and the remaining two are assigned 0.

$$\widehat{V}_{jnm} = -p_{jnm} + \widehat{ASC}_{jn} + \beta_{feat} * feat_{jnm}, j = 1, 2, 3, 4 \tag{42}$$

As noted in section 2, if we carry out estimation using the two datasets combined, the estimates will be consistent. However, if we carry out estimation using the endogenous dataset only, we observe substantial bias as shown in Table 7. The parameter of display has a positive bias, and the scale parameter has a negative bias.

4.2.3. Control-function correction

As in the previous example, we can apply the CF correction described in Section 3.2 to address endogeneity. However, in this example, the only endogenous variables are $disp_{jnm}$. The instruments can be obtained using Equation (32), but substituting the estimated household-specific parameters $\widehat{ASC}_{jn}$ by the estimated population parameters $\widehat{\widehat{ASC}}_j$, in order to obtain *what would have been displayed to an average household*. As explained before, these instruments are correlated with the personalized display variable (because both are partially determined by the price and feature attributes), but not with the estimated household-specific parameters. The residuals of the four display attributes (for each alternative) are then obtained from probit regressions where the display variables (for

Table 6
Posterior means for the catsup dataset (standard deviations are shown in parentheses).

	ASC - Heinz41	ASC - Heinz32	ASC - Heinz28	Display	Featured	Scale
Population means	1.236	1.184	1.951	0.583	0.651	0.681
Population variances	1.665	1.492	1.362	0.243	(No significant heterogeneity)	

Table 7
Posterior means using the endogenous dataset using the uncorrected model and the CF correction (standard deviations in parentheses).

	ASC - Heinz41	ASC - Heinz32	ASC - Heinz28	Display	Featured	Scale
Population Means						
True values	1.236	1.184	1.951	0.583	0.651	0.681
CF Correction	1.213 (0.028)	1.130 (0.024)	1.945 (0.026)	0.692 (0.036)	0.660 (0.026)	0.652 (0.018)
Uncorrected	1.331 (0.044)	1.233 (0.036)	2.011 (0.038)	1.813 (0.059)	0.624 (0.034)	0.298 (0.022)
Variances (imputed for the CF correction)						
True values	1.665	1.492	1.362	0.243	No heterogeneity	
CF Correction	1.511 (0.056)	1.373 (0.039)	1.340 (0.042)	0.424 (0.038)		
Uncorrected	2.535 (0.157)	2.043 (0.116)	1.844 (0.121)	0.737 (0.074)		

each alternative) are the dependent variables, and the instruments are the independent variables. The vector of residuals, referred to as $\widehat{\mu}_n$, constitutes of the four residuals $\{\mu_{Heinz41}, \mu_{Heinz32}, \mu_{Heinz28}, \mu_{Hunts32}\}$ averaged over the household's choice situations. The modified utility equations are shown in Equation (43).

$$V_{jnm} = \left(-p_{jnm} + ASC_{jn} + \lambda_j \widehat{\mu}_n + \left(\beta_{disp,n} + \widehat{\mu}_n\right) * disp_{jnm} + \beta_{feat} * feat_{jnm}\right) * \exp\left(\alpha\right) \tag{43}$$

As explained in Section 3.2, the control function term is not only included as a separate term in the utility equation, but also interacted with the display attribute, $disp_{jnm}$.

The model is estimated using the same Gibbs sampler explained in Section 3.2.2, and convergence is checked using the Gelman and Rubin's convergence diagnostic, and the Heidelberger and Welch's convergence diagnostic, which are available in the R package 'coda'.

Similar to the above example, we used two million draws, from which we discarded the first one million, and saved every 10,000th draw from the remaining million. The estimation results shown in Table 7 indicate that the CF correction reduces the bias substantially, but does not completely eliminate it. In the following section, we assess the performance of this method to different sample sizes and panel lengths (choice histories).

4.2.4. *Varying sample and history size*

In the above example, we used a sample size of 5000 individuals, and a history of 5 choices per individual. In this section, we

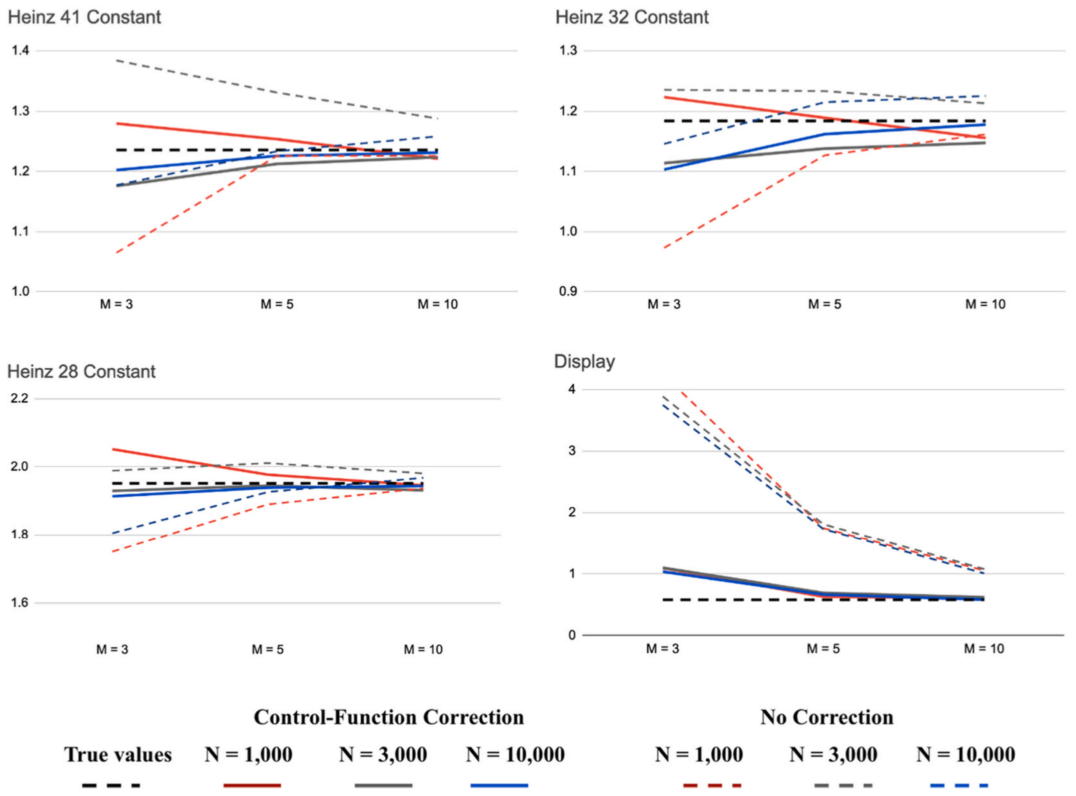

Fig. 3. Sensitivity analysis with respect to sample and history size: posterior means of the population means.

consider additional samples of 3000 and 10,000 individuals, and for each of these samples, we consider a choice history of 3, 5, or 10 choices.

The results of these estimations are shown in Figs. 3–5. These results indicate that endogeneity introduces significant bias not only in the coefficient of the endogenous variable, but in most of the other population means, variances, and in the scale parameter. They also confirm the results of Danaf et al. (2020), which showed that bias tends to decrease with longer choice histories.

The control-function correction reduces the magnitude of bias substantially across most parameters. However, bias is still present even after correction, especially with very short histories (M = 3). With 10 choices per individual, the magnitude of bias is negligible and not statistically significant for most of the estimated parameters. These results are also consistent with the findings of Orme (2001), who found that a control-function correction reduced bias substantially in dynamic non-linear panel data models, but did not completely eliminate it.

These results are also consistent with the findings of Akay (2012), who found that the Wooldridge method works very well for panels of moderately long duration (longer than 5–8 periods) in dynamic panel data models. The latter study also concluded that the performance of the Wooldridge method for short panels may be highly sensitive to the specification of the auxiliary distribution of the unobserved individual-effects and the explanatory variables entering into the specification, which in our case, are the estimated residuals.

In both examples discussed in Sections 4.1 and 4.2, the CF correction reduces the bias substantially. This is also the case even when the estimated individual/household-specific parameters are obtained from a misspecified model (for example, if these are estimated using an endogenous dataset instead of an exogenous one), as long as the correlation between the instruments and the endogenous variables is maintained.

5. Discussion

The method presented in Section 3.2 corrects for endogeneity bias resulting from the correlation between the explanatory variables and the individual-specific effects. Recommendation systems and personalized advertisement are common examples of the latter case because individuals with a higher preference for a specific attribute are more likely to be presented with alternatives having higher values of this attribute. The CF correction described in Section 3.2 can apply to various contexts when relevant instruments are available. For example, in travel mode choice, individuals with a higher time sensitivity might choose to live near their workplace, resulting in lower travel times.

Most of the control-function applications in discrete choice models are not suitable for addressing this issue, because they consider only correlations between one or more attributes, and the unobserved error term ε_{jmn}. However, it is not sufficient to include the residuals in the utility equations in such cases; both the residuals and their interactions with the attributes should be included.

5.1. Consistency of the estimates

The proposed control-function method reduces the observed bias in the parameter estimates substantially. However, according to Wooldridge (2015), the consistency of the CF estimator depends on two assumptions: the first stage being correctly specified (i.e., Equation (18)), and the linearity assumptions in Equation (21).

In the above examples, we used a linear specification in the first stage for continuous attributes, and a probit model for binary attributes. However, more flexible specifications can be adapted in order to exploit the information in the exogenous instruments (including non-linearities) while avoiding overfitting, such as random forests. Future research should look into flexible specifications of the first stage that improve the consistency of the control-function estimator.

Finally, in the above applications, we used the attributes of non-personalized recommendations (e.g., recommendations to an

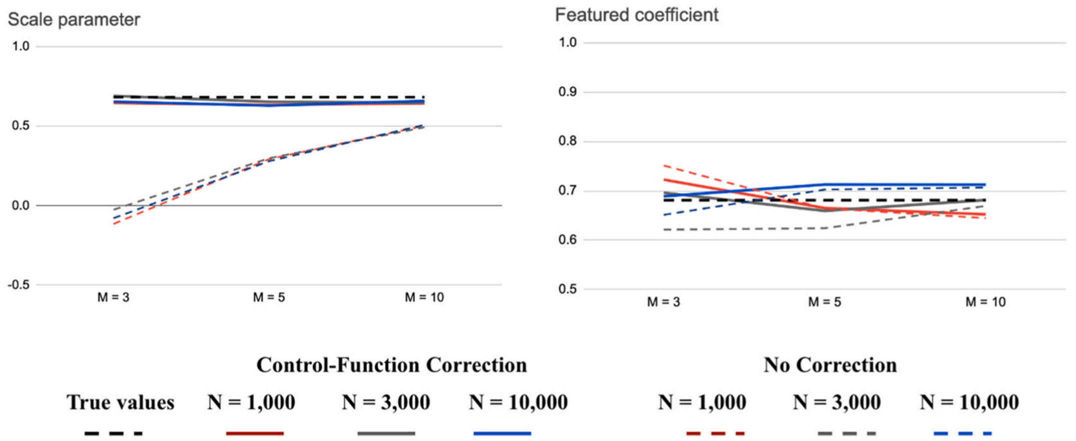

Fig. 4. Sensitivity analysis with respect to sample and history size: posterior means of the fixed parameters.

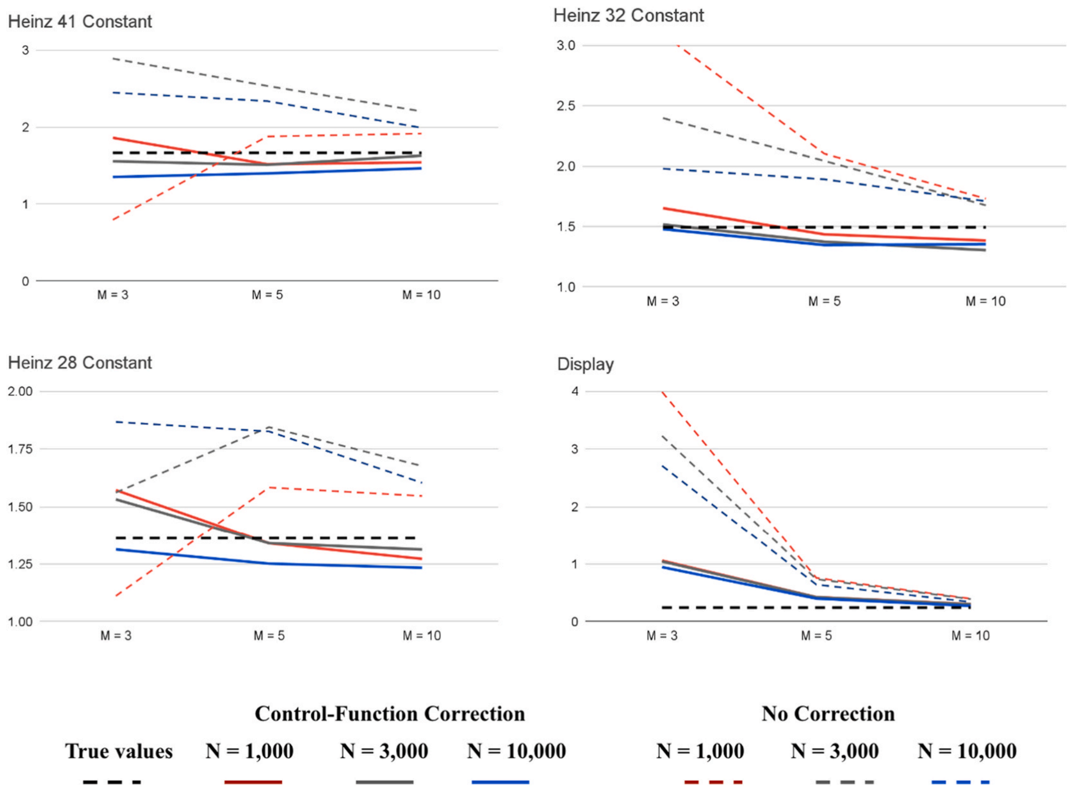

Fig. 5. Sensitivity analysis with respect to sample and history size: posterior means of the population variances.

average individual) as instruments. This is relevant for this specific example because these instruments are correlated with the attributes of the recommended alternatives (both recommended sets are generated from the same inventory), and uncorrelated with the individual-specific effects. However, in some cases, non-personalized recommendations might result in weak instruments, because of large heterogeneity, or large cardinality of the inventory. Future work should look into extensions of statistical tests of relevance and exogeneity, to the context discussed in this paper.

5.2. *Simultaneous estimation*

The CF method described above can be estimated simultaneously in order to recover the correct standard errors directly (from the inverse of the information matrix in frequentist estimation, or as the posterior standard deviations in Bayesian estimation). Sequential estimation usually underestimates the standard errors because the residuals used in the second stage are estimated in the first stage (Guevara and Hess, 2019; Villas-Boas and Winer, 1999).

Bootstrapping and the delta method can be used in order to estimate the standard errors with sequential estimation. However, Bootstrapping is computationally expensive especially in logit mixture models (which require simulation using HB or Maximum Simulated Likelihood (MSL)). In addition, simultaneous estimates would still be more efficient (smaller standard errors), assuming the model captures the true data generation process, because they can attain the Cramér–Rao lower bound.

We can use the approach proposed by Villas-Boas and Winer (1999) and Park and Gupta (2009) which is the simultaneous estimation of the first stage and the choice model. In Bayesian estimation, the first stage regressions are included in the Gibbs sampler and residuals are updated at each iteration.

5.3. *Alternative methods*

In the context of recommendation systems, the CF method described in this paper is only relevant when the historical data cannot be included in estimation. As demonstrated by Danaf et al. (2020), correction is not needed when the entire choice history of each individual can be used. However, in such systems, the number of individuals and choices per individual can be large, resulting in computational constraints. In addition, Danaf et al. (2020) show that with large enough panels (i.e., number of observations per individual), the bias diminishes to negligible levels.

An alternative method would be to use a full information maximum likelihood estimator, by modeling the choice probability and recommendation probability jointly. This involves approximating the expression in Equation (8), in order to model the probability of recommending attributes X_m conditional on the individual-specific parameters β_n (for example, we can use a choice model representing the recommender system's choice of recommendations from the initial inventory). Ongoing research is focusing on exploring alternative correction methods and comparing them to CF estimators.

6. Conclusion

This paper presented a control-function correction for endogeneity that arises in random coefficients models as a result of correlation between the explanatory variables and the individual-specific parameters. This method extends the standard CF method by including interaction terms between the first stage residuals and the explanatory variables.

In such cases, relevant instruments need to be correlated with the explanatory variables, but not with the individual-specific coefficients. In our application to a choice-based recommender system, we used non-personalized recommendations as instruments. This can be extended to other contexts. For example, in endogenous SP designs (e.g., SP profiles that are constructed based on RP data), the instruments can either be the RP attributes (as proposed by Guevara and Hess (2019)), or profile attributes that are generated to an average individual for a similar choice situation.

While this method results in good estimates of the population means, the original covariance matrix is not consistently estimated. The estimated variances are deflated because part of the inter-consumer heterogeneity is explained by the residuals. However, this is harmless because we can include the residuals in forecasting and model application. Our simulation results show that using residuals in model application can even outperform the predictions of the true parameters. In addition, we can recover the true variances from the estimation results as demonstrated in Section 4.1.5.

One of the main limitations of this method is that it assumes that the parameters are normally or log-normally distributed. Future research should focus on extending the CF method or proposing new methods that can deal with other mixing distributions. Another limitation of this method is that in some cases, it might require the estimation of too many auxiliary parameters. In the Monte Carlo experiment presented in Section 4.1, 60 additional parameters were estimated even though there were only four endogenous variables. In some cases, this might require larger sample sizes, and result in longer estimation times.

In ongoing research, Xie et al. (2022) are applying a similar control-function correction to the case of dynamic choice models with preference heterogeneity (where the utility of an alternative at a given time period depends on the individual's choice at the previous time period), as an alternative to the Wooldridge (2005) and Heckman (1978) corrections. The CF method and the Wooldridge method exhibit some similarities when there is a single continuous endogenous variable; in both cases, the random parameters are expressed as a function of exogenous variables or instruments derived from these exogenous variables. Our preliminary results in Monte Carlo experiments indicate that both methods result in similar parameter estimates with reduced bias (compared to the uncorrected model). However, the CF method results in smaller standard errors.

Author statement

Mazen Danaf, Angelo Guevara, and Moshe Ben-Akiva conceived of the presented idea and developed the theory. Mazen Danaf performed the experiments and computations. Angelo Guevara and Moshe Ben-Akiva supervised the findings of this work. All authors discussed the results and contributed to the final manuscript. Mazen Danaf wrote the manuscript with support from Angelo Guevara and Moshe Ben-Akiva.

Acknowledgements

This work is partially aligned with the TRIPOD: Sustainable Travel Incentives with Prediction, Optimization and Personalization research project sponsored by the U.S. Department of Energy Advanced Research Projects Agency-Energy (ARPA-E). It was awarded through the ARPA-E Traveler Response Architecture using Novel Signaling for Network Efficiency in Transportation (TRANSNET) program. Angelo Guevara was funded in part by CONICYT, FONDECYT 1191104 and by ANID PIA/PUENTE AFB220003.

References

Abernethy, J., Evgeniou, T., Toubia, O., Vert, J.P., 2007. Eliciting consumer preferences using robust adaptive choice questionnaires. IEEE Trans. Knowl. Data Eng. 20 (2), 145–155.

Akay, A., 2012. Finite-sample comparison of alternative methods for estimating dynamic panel data models. J. Appl. Econom. 27 (7), 1189–1204.

Ben-Akiva, M., McFadden, D., Train, K., 2019. Foundations of stated preference elicitation: consumer behavior and choice-based conjoint analysis. Found. Trend. Econom. 10 (1–2), 1–144.

Berry, S., Levinsohn, J., Pakes, A., 1995. Automobile prices in market equilibrium. Econometrica: J. Econom. Soc. 841–890.

Bhat, C.R., 2000. Incorporating observed and unobserved heterogeneity in urban work travel mode choice modeling. Transport. Sci. 34 (2), 228–238.

Blundell, R., Powell, J.L., 2003. Endogeneity in nonparametric and semiparametric regression models. Econom. Soc. Monogr. 36, 312–357.

Bradley, M.A., Daly, A.J., 1993. New Analysis Issues in Stated Preference Research. PTRC-PUBLICATIONS-P, 75-75.

Card, D., 1993. Using Geographic Variation in College Proximity to Estimate the Return to Schooling (No. W4483). National Bureau of Economic Research.

Chaptini, B.H., 2005. Doctoral dissertation Use of discrete choice models with recommender systems. Massachusetts Institute of Technology.

Croissant, E., 2014. Mlogit: Multinomial Logit Models. R library.

Danaf, M., Becker, F., Song, X., Atasoy, B., Ben-Akiva, M., 2019. Online discrete choice models: applications in personalized recommendations. Decis. Support Syst. 119, 35–45.

Danaf, M., Atasoy, B., Ben-Akiva, M., 2019b. Observed and unobserved inter- and intra-consumer heterogeneity. In: Accepted for Presentation in the 6th International Choice Modelling Conference (ICMC) (Kobe, Japan).

Danaf, M., Guevara, A., Atasoy, B., Ben-Akiva, M., 2020. Endogeneity in adaptive choice contexts: choice-based recommender systems and adaptive stated preferences surveys. J. Choice Modell. (in press).

Fowkes, A.S., Shinghal, N., 2002. The Leeds Adaptive Stated Preference Methodology.

Fowkes, T., 2007. The design and interpretation of freight stated preference experiments seeking to elicit behavioural valuations of journey attributes. Transp. Res. Part B Methodol. 41 (9), 966–980.

Frazier, D.T., Renault, E., Zhang, L., Zhao, X., 2020. Weak Identification in Discrete Choice Models arXiv preprint arXiv:2011.06753.

Garen, J., 1984. The returns to schooling: a selectivity bias approach with a continuous choice variable. Econometrica 52 (5), 1199.

Guevara, C.A., 2018. Overidentification tests for the exogeneity of instruments in discrete choice models. Transp. Res. Part B Methodol. 114, 241–253.

Guevara, C.A., Ben-Akiva, M., 2006. Endogeneity in residential location choice models. Transport. Res. Rec. 1977 (1), 60–66.

Guevara, C.A., Ben-Akiva, M., 2010. Addressing endogeneity in discrete choice models: assessing control-function and latent-variable methods. In: Choice Modelling: the State-Of-The-Art and the State-Of-Practice: Proceedings from the Inaugural International Choice Modelling Conference. Emerald Group Publishing Limited, pp. 353–370.

Guevara, C.A., Ben-Akiva, M.E., 2012. Change of scale and forecasting with the control-function method in logit models. Transport. Sci. 46 (3), 425–437.

Guevara, C.A., Hess, S., 2019. A control-function approach to correct for endogeneity in discrete choice models estimated on SP-off-RP data and contrasts with an earlier FIML approach by Train & Wilson. Transp. Res. Part B Methodol. 123, 224–239.

Guevara, C.A., Polanco, D., 2016. Correcting for endogeneity due to omitted attributes in discrete-choice models: the multiple indicator solution. Transportmetrica: Transport. Sci. 12 (5), 458–478.

Guevara, C.A., 2015. Critical assessment of five methods to correct for endogeneity in discrete-choice models. Transport. Res. Pol. Pract. 82, 240–254.

Hausman, J.A., 1978. Specification tests in econometrics. Econometrica: J. Econom. Soc. 1251–1271.

Heckman, J.J., 1978. Dummy Endogenous Variables in a Simultaneous Equation System. National Bureau of Economic Research. Technical report.

Horsky, D., Misra, S., Nelson, P., 2006. Observed and unobserved preference heterogeneity in brand-choice models. Market. Sci. 25 (4), 322–335.

Huang, A., Wand, M.P., 2013. Simple marginally noninformative prior distributions for covariance matrices. Bayesian Anal. 8 (2), 439–452.

Huber, J., Train, K., 2001. On the similarity of classical and Bayesian estimates of individual mean partworths. Market. Lett. 12 (3), 259–269.

Jain, D.C., Vilcassim, N.J., Chintagunta, P.K., 1994. A random-coefficients logit brand-choice model applied to panel data. J. Bus. Econ. Stat. 12 (3), 317–328.

Jiang, H., Qi, X., Sun, H., 2014. Choice-based recommender systems: a unified approach to achieving relevancy and diversity. Oper. Res. 62 (5), 973–993.

Johnson, F.R., Kanninen, B., Bingham, M., Özdemir, S., 2006. Experimental design for stated-choice studies. In: Valuing Environmental Amenities Using Stated Choice Studies. Springer, Dordrecht, pp. 159–202.

Johnson, F.R., Lancsar, E., Marshall, D., Kilambi, V., Mühlbacher, A., Regier, D.A., Bresnahan, B.W., Kanninen, B., Bridges, J.F., 2013. Constructing experimental designs for discrete-choice experiments: report of the ISPOR conjoint analysis experimental design good research practices task force. Value Health 16 (1), 3–13.

Kanninen, B.J., 2002. Optimal design for multinomial choice experiments. J. Market. Res. 39 (2), 214–227.

Koop, G., Poirier, D.J., Tobias, J.L., 2007. Bayesian Econometric Methods. Cambridge University Press.

Orme, C.D., 2001. Two-step Inference in Dynamic Non-linear Panel Data Models. University of Manchester. Unpublished paper. http://personalpages.manchester.ac.uk/staff/chris.orme/documents/Research%20Papers/initcondlast.pdf.

Park, S., Gupta, S., 2009. Simulated maximum likelihood estimator for the random coefficient logit model using aggregate data. J. Market. Res. 46 (4), 531–542.

Petrin, A., Train, K., 2010. A control function approach to endogeneity in consumer choice models. J. Market. Res. 47 (1), 3–13.

Plummer, M., Best, N., Cowles, K., Vines, K., Sarkar, D., Bates, D., Almond, R., Magnusson, A., 2020. R Package 'coda': Output Analysis and Diagnostics for MCMC. Retrieved from. https://cran.r-project.org/web/packages/coda/coda.pdf.

Revelt, D., Train, K., 2000. Customer-specific Taste Parameters and Mixed Logit: Households' Choice of Electricity Supplier. Department of Economics, UC Berkeley.

Ricci, F., Rokach, L., Shapira, B., 2015. Recommender systems: introduction and challenges. In: Recommender Systems Handbook. Springer, Boston, MA, pp. 1–34.

Shinghal, N., 1999. An Application of Stated Preference Methods to the Study of Intermodal Freight Transport Services in India. Institute for Transport Studies University of Leeds. Phd Thesis (unpublished), October 1999.

Song, X., Atasoy, B., Ben-Akiva, M., 2017. Smart Mobility through Personalized Menu Optimization (No. 17-05906.

Song, X., Danaf, M., Atasoy, B., Ben-Akiva, M., 2018. Personalized menu optimization with preference updater: a Boston case study. Transport. Res. Rec. 2672 (8), 599–607.

Teo, C.H., Nassif, H., Hill, D., Srinivasan, S., Goodman, M., Mohan, V., Vishwanathan, S.V.N., 2016, September. Adaptive, personalized diversity for visual discovery. In: *Proceedings of the 10th ACM conference on recommender systems*, pp. 35–38.

Toubia, O., Simester, D.I., Hauser, J.R., Dahan, E., 2003. Fast polyhedral adaptive conjoint estimation. Market. Sci. 22 (3), 273–303.

Train, K., Weeks, M., 2005. Discrete choice models in preference space and willingness-to-pay space. In: Applications of Simulation Methods in Environmental and Resource Economics. Springer, Dordrecht, pp. 1–16.

Train, K., Wilson, W.W., 2008. Estimation on stated-preference experiments constructed from revealed-preference choices. Transp. Res. Part B Methodol. 42 (3), 191–203.

Train, K.E., Wilson, W.W., 2009. Monte Carlo analysis of SP-off-RP data. J. Choice Modell. 2 (1), 101–117.

Train, K.E., 2009. Discrete Choice Methods with Simulation. Cambridge university press.

Villas-Boas, J.M., Winer, R.S., 1999. Endogeneity in brand choice models. Manag. Sci. 45 (10), 1324–1338.

Wooldridge, J.M., 2005. Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. J. Appl. Econom. 20 (1), 39–54.

Wooldridge, J.M., 2015. Control function methods in applied econometrics. J. Hum. Resour. 50 (2), 420–445.

Xie, Y., Danaf, M., Guevara, A., Ben-Akiva, M., 2022. A control function solution to the initial conditions problem in dynamic choice models with random parameters. In: Working Paper. Massachusetts Institute of Technology.

Zhu, X., Wang, F., Chen, C., Reed, D.D., 2019. Personalized Incentives for Promoting Sustainable Travel Behaviors. Transportation Research Part C: Emerging Technologies.