Intern. J. of Research in Marketing 31 (2014) 35–48

Contents lists available at ScienceDirect

Intern. J. of Research in Marketing

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

A simple method for estimating preference parameters for individuals

Bart D. Frischknecht [a,*], Christine Eckert [a,b], John Geweke [a,c], Jordan J. Louviere [a,b]

[a] *Centre for the Study of Choice, University of Technology, Sydney, Australia*
[b] *Marketing Discipline Group, University of Technology, Sydney, Australia*
[c] *Economics Discipline Group, University of Technology, Sydney, Australia*

ARTICLE INFO

Article history:
First received in 9 March 2012 and was under review for 11 months
Available online 1 October 2013

Area Editor: Bruce G.S. Hardie

Keywords:
Choice modeling
Bayesian estimation
Discrete choice experiment
Conjugate prior
Individual level choice model

ABSTRACT

This paper demonstrates a method for estimating logit choice models for small sample data, including single individuals, that is computationally simpler and relies on weaker prior distributional assumptions compared to hierarchical Bayes estimation. Using Monte Carlo simulations and online discrete choice experiments, we show how this method is particularly well suited to estimating values of choice model parameters from small sample choice data, thus opening this area to the application of choice modeling. For larger sample sizes of approximately 100–200 respondents, preference distribution recovery is similar to hierarchical Bayes estimation of mixed logit models for the examples we demonstrate. We discuss three approaches for specifying the conjugate priors required for the method: specifying priors based on existing or projected market shares of products, specifying a flat prior on the choice alternatives in a discrete choice experiment, or adopting an empirical Bayes approach where the prior choice probabilities are taken to be the average choice probabilities observed in a discrete choice experiment. We show that for small sample data, the relative weighting of the prior during estimation is an important consideration, and we present an automated method for selecting the weight based on a predictive scoring rule.

1. Introduction

Populations of individuals are heterogeneous in their choices. There are many ways to model heterogeneity in quantitative models of human choice. Estimating a different choice model for each person is a conceptually attractive approach, but it is often thwarted by data separation. The investigator is then left with the more demanding procedure of pooling the data across individuals and setting up a formal distribution of preferences in a population. In the common paradigm of heterogeneous choice model estimation, the investigator specifies a prior over the preference space (Allenby, Arora, & Ginter, 1995; Rossi, Allenby, & McCulloch, 2005). The specification of a prior over the preference space requires the specification of the distributional form of the prior as well as parameter values for the distribution. The approach proposed in this article overcomes the issue of data separation without requiring a prior distribution on the preference parameters or pooling of the data across respondents. The method is based on maximum likelihood estimation of a logit model and thus provides a simple way to estimate a choice model for a single individual. Applied over a sample of individuals, the proposed method shows a similar performance in recovering the sample distribution of preferences and reduced computational complexity compared to hierarchical Bayes (HB) estimation.

1.1. Data separation

A critical drawback in estimating a different choice model for each person by maximum likelihood is that a single individual's data often exhibit data separation, whereby the responses of the individual can be perfectly classified by a linear combination of the covariates described in Albert and Anderson (1984), and updated by Santner and Duffy (1986). Complete separation occurs when a combination of explanatory variables classifies responses without error according to a strict inequality. Quasi-complete separation occurs when a combination of explanatory variables classifies responses without error up to a non-strict inequality. All other cases are considered to exhibit data overlap. Cases of complete or quasi-complete separation are more likely in small samples or when a particular alternative is chosen with low probability, which has been recognized in biostatistics literature with application to clinical trials (Heinze, 2006) and in econometrics (Beggs, Cardell, & Hausman, 1981) and marketing (Chapman, 1984) literature with previous efforts to estimate choice models using a small sample of data from a single individual.[1]

[*] Corresponding author.

[1] Encountering data separation in choice data can be interpreted in two ways. One interpretation is that the underlying choice behavior is stochastic and that multinomial logistic regression is a suitable model to describe the observed choice behavior. Here, data separation is an artifact of a relatively small number of observations. The second interpretation is that the data separation is evidence of a deterministic choice process such as lexicographic decision-making. The second interpretation would indicate that multinomial logistic regression is inappropriate for classifying the data at hand. In that case, we expect data separation to persist as the number of observations increases. In this article, we adopt the first interpretation.

0167-8116/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ijresmar.2013.07.005

In the case of complete or quasi-complete data separation (Albert & Anderson, 1984), maximum likelihood estimates for multinomial logistic regression (Train, 2003), one of the most commonly applied choice models in marketing and economics, do not exist. Maximum likelihood estimation in these cases implies that the estimates of the parameters are unbounded. Data overlap alone does not guarantee a sufficiently small bias that would result in satisfactory estimates, as shown by King and Ryan (2002), who investigated a case of near separation in which data overlap existed but was relatively small.

1.2. Maximum penalized likelihood estimation

Although the complete separation problem has been encountered previously in econometrics and marketing (Beggs et al., 1981; Chapman, 1984; Savin & Wurtz, 1999), it has likely attracted limited attention because of the extensive use of data pooling across respondents, leading to large samples and data overlap rather than separation. Various means have been proposed to overcome the data separation challenge, especially for biostatistics applications, due to the small sample sizes and low incidence rates of many clinical trials (Bull, Mak, & Greenwood, 2002; Cardell, 1993; Clogg, Rubin, Schenker, Schultz, & Weidman, 1991; Firth, 1993; Heinze, 2006). The proposed approaches, including the approach used in this article, rely on the principle of shrinkage as described by Stein (1956) and James and Stein (1961).

In our case, the shrinkage is accomplished by estimation of parameters based on the maximum penalized likelihood. The penalty function that we adopt is to augment the limited data with prior beliefs about the behavior of the data (Geweke, 2005), which corresponds to a Bayesian approach designed to overcome the challenge of using finite samples. The relevance of a penalty approach to econometric or marketing choice problems as a method for capturing sample heterogeneity appears not to have been recognized by the choice modeling community (see, for example, section 1 of Allenby & Rossi, 1999) until Evgeniou, Pontil, and Toubia (2007).

To date, maximum penalized likelihood approaches can be classified as either fixed penalty methods or updated penalty methods, according to the penalty function adopted.[2] We first discuss fixed penalty methods, to which the approach used in this paper belongs. Fixed penalty methods add a carefully considered fixed set of artificial observations to the data, thereby ensuring data overlap for the extended sample. Haldane (1955), motivated by reducing parameter estimate bias in a binomial logit case, suggests a change to the likelihood formulation for an estimation that adds an artificial observation to the data for each binary outcome. Each artificial observation is given half the weight of one of the original observations in the log likelihood function. Both Clogg et al. (1991) and Cardell (1993), motivated by data separation (see also Beggs et al., 1981), propose artificially generating sets of observations (or chosen and unchosen alternatives) coupled with specific explanatory variables that are generated in a particular way. Clogg et al. (1991) illustrate their approach only for a binomial case. They consider the relative outcome frequency observed in the data and the number of estimation parameters to determine the number of artificial observations.

In general, the fixed penalty methods that add artificial observations to the data are examples of applying a conjugate prior to the data, i.e., a prior that has the same distributional form as the likelihood function. We discuss the priors employed in this paper later in the manuscript. The Cardell (1993) approach can be applied to a binomial or multinomial case and is intended to be applied to choice data rather than clinical trials or census demographics. It adds J artificial choice task observations where J is the total number of unique outcome alternatives (e.g., car, bus, train). The chosen alternative in each artificial choice task is represented by the average of the explanatory variables associated with the alternative from the choice tasks when the alternative was not chosen in the original data set. Overlap is ensured in this way by adding artificial observations that are opposite to the observed data. Because the artificial observations are composed based on the design and the choice responses for a particular alternative, the Cardell (1993) approach appears most appropriate for alternative specific choice models. Even in this case, the interpretation of the artificial observations as a conjugate prior is dependent on the specific values of the explanatory variables.

A more complex alternative to fixed penalty methods is to derive an updated penalty, which is a penalty function that is a function of the estimated model itself. Firth (1993), initially motivated by the goal of reducing parameter estimate bias, illustrates an approach for updating the penalty function at each iteration of a numerical procedure for maximizing the log likelihood function. Heinze and Schemper (2002) for binary and Bull et al. (2002) for multinomial logistic regression recognize that Firth's technique can be applied to the case of separated data and expand on his approach. Evgeniou et al. (2007) develop an updating penalty method for maximum penalized likelihood estimation and applies this method to discrete choice data. Gilbride, Lenk, and Brazell (2008) and Evgeniou et al. (2007) find that this method produces point estimates and predictions very similar to those of hierarchical Bayes estimation.

1.3. Proposed approach

For reasons we discuss in more detail below, we propose the use of an approach in the tradition of fixed penalty methods. This approach is similar to those of Clogg et al. (1991) and Cardell (1993), except Clogg et al. (1991) present their method only for a binomial case and for repeated observations of the vectors of explanatory variables, and the prior employed in Cardell's method is not interpretable for generic, i.e., unlabeled-alternative, choice models.

We present our method for a multinomial case in choice model format and for explanatory variables that vary between alternatives, such that our approach is readily applied to data collected from a single individual completing an unlabeled discrete choice experiment. This article further differs from similar methods presented previously in economics and statistics literature that focus on data sets of sufficient size in the following ways: the relative weight given to the prior during estimation does not have a large impact, and the methods present a specific prior weighting strategy without regard to its effect on the prediction performance of the estimated model. We show that for small sample data, the relative weighting of the prior during estimation is an important consideration, and we present an automated method for selecting a prior weight based on a predictive scoring rule.

We discuss three ways to formulate a conjugate prior for application to data from discrete choice experiments. First, the investigator can specify beliefs about the choice shares of the alternatives presented in the discrete choice experiment. For example, a flat prior, or probability $\pi_j = 1 / J$, for $j = 1,\ldots,J$ alternatives shrinks parameter estimates towards zero, implying that each alternative is equally likely. It thus pulls the maximum likelihood estimates away from $\pm\infty$ in the case of separated data. Similarly, an alternative to a flat prior for alternatives in the discrete choice experiment is to adopt an empirical Bayes approach (Carlin & Louis, 2000), where the prior implies the aggregate choice shares of the discrete choice experiment alternatives observed in the sample population. Finally, given a specific set of product alternatives available in the market, an investigator can specify the observed market share of each alternative or her beliefs about the relative market share of each alternative, as in the case of a new product introduction. These three approaches are illustrated in subsequent examples.

[2] Apart from fixed and updated penalty methods, exact logistic regression (Mehta & Patel, 1995) has been proposed as an alternative to maximum penalized likelihood estimation when data are separated. However, its application is limited in many practical cases because the method is computationally more intense than the maximum penalized likelihood, continuous explanatory variables are not handled well and confidence intervals can be overly conservative (Heinze, 2006). Additionally, Heinze and Schemper (2002) and Heinze (2006) compare exact logistic and penalized likelihood approaches for logistic regression with separated data and conclude that the penalty method is superior in most instances.

While we leave a performance comparison between our fixed penalty approach and the updating penalty approach of Evgeniou et al. (2007) for future study, our approach has three notable advantages for the market research community. First, our method allows investigators to directly specify priors in terms of choice shares rather than parameter distributions. Second, the weight given to the penalty component of the penalized likelihood function intuitively represents the relative contribution to the model estimation of the prior belief compared to the observed data. Third, the fixed penalty can be implemented using standard software for maximum likelihood estimation.

The performance of our proposed method is demonstrated across a range of sampling conditions, including Monte Carlo simulations and real-world data sets from discrete choice experiments. The results suggest that it is feasible and convenient to estimate a discrete choice model for small samples, including a single individual using the modified maximum likelihood approach and data collected from a discrete choice experiment.

The remainder of the article is organized as follows. Section 2 develops the proposed method for specifying a conjugate prior based on choice shares that overcomes the complete separation problem in multinomial logistic regression with small samples. Section 3 provides a series of Monte Carlo studies that explore the effect of the proposed method on discrete choice model parameter estimation and prediction. Section 4 illustrates the application of the approach to four data sets from online discrete choice experiments, including a comparison to HB estimation of a mixed logit model. Section 5 discusses the influence of the design of the choice experiment on the proposed estimation method. Section 6 concludes the article.

2. Method

We explain our modified maximum likelihood method (MML) in the context of a data set collected from a discrete choice experiment (Louviere, Hensher, & Swait, 2000). Similar to Clogg et al. (1991), we specify a conjugate prior so that the posterior for the choice model parameters β has the same form as the likelihood function, thus enabling the methodology to be implemented using standard software. The conjugate prior described below is implemented through the construction of a notional, or imaginary, sample of data.

The artificial data are called a notional sample because the data take the same form as they would if a fictitious respondent provided responses to fictitious choice tasks. In our example, the fictitious choice tasks are chosen to be the same choice tasks faced by the respondents in the discrete choice experiment or the same choice tasks faced by the respondents in the discrete choice experiment and some additional choice tasks chosen for their similarity to the market of interest. The fictitious responses are determined in one of three ways, as described below, and correspond to three different interpretations of the prior.

First, the investigator can assign equal probabilities to all alternatives in all choice tasks. Second, when the notional sample choice tasks coincide with the discrete choice experiment choice tasks, the investigator can adopt an empirical Bayes approach and assign choice shares based on the frequency that each alternative is chosen by the sample of respondents. Third, the investigator can consider each choice task of the notional sample and assign choice shares (i.e., probabilities) to each alternative based on subjective belief or external market data in the case of market realistic choice tasks.

For clarity of exposition in Section 2.1, we begin with a notional sample that consists of the same choice tasks observed by respondents in a discrete choice experiment, and we adopt a prior belief that all alternatives are equally likely to be chosen, or probability $\pi_j = 1 / J$, for $j = 1,\ldots,J$ alternatives. In Section 2.2, we demonstrate two additional approaches for forming a notional sample: the empirical Bayes approach and the combination of the flat prior from Section 2.1 with additional market relevant choice tasks.

The method is described in the context of the data collected from a single individual. We estimate a separate choice model for each individual in the sample using the modified maximum likelihood method. For any post-hoc simulation studies, we then assume that the distribution of sample preferences represents the distribution of population preferences or that the sample can be weighted in such a way that the resulting distribution of preferences represents the distribution of population preferences.

2.1. Mathematical description assuming a flat prior

Assume that we have a sample of S completed choice tasks, where each choice task includes J alternatives[3] and each alternative is described by a $k \times 1$ vector of explanatory variables x_{sj}, where alternatives $j = 1,\ldots,J$ may or may not be related across choice tasks, i.e., a labeled or unlabeled design. The dependent variable $y_{sj} = 1$ when alternative j is chosen in choice task s, and $y_{sj} = 0$ otherwise.

The conventional likelihood function for multinomial logistic regression is given by

$$L = \prod_{s=1}^{S}\prod_{j=1}^{J}\pi_{sj}^{y_{sj}} \tag{1}$$

where

$$\pi_{sj} = \frac{e^{\beta' x_{sj}}}{\sum_{l=1}^{J} e^{\beta' x_{sl}}}. \tag{2}$$

The conjugate prior is given by

$$p(\beta) \propto \prod_{s=1}^{S}\prod_{j=1}^{J}\pi_{sj}^{\alpha g_{sj}}, \tag{3}$$

where α is a positive constant to be specified and g_{sj} are weight proportions that conceptually correspond to the choice shares of each alternative in the notional sample choice tasks. The posterior is then

$$p(\beta|y) \propto \prod_{s=1}^{S}\prod_{j=1}^{J}\pi_{sj}^{y_{sj}+\alpha g_{sj}}. \tag{4}$$

To illustrate, when the weight proportions are constrained to be evenly divided as $g_{sj} = 1 / J$, then at the mode of the prior distribution, all choices are equally likely for all points $\mathbf{x}_{sj}$ in the sample. As α becomes larger, the prior information that choices are equally likely at all sample points $\mathbf{x}_{sj}$ becomes stronger. Given that there are S original observations each with a weight y_{sj} of unity and JS artificial observations, each with a weight of $ag_{sj} = \alpha / J$, the ratio R_{prior} of the artificial observations, i.e., the prior as described by a notional sample, to the original data observations is

$$R_{prior} = \frac{\alpha S}{S} = \alpha. \tag{5}$$

Therefore, values of $\alpha > 1$ result in greater weight on the notional prior than on the observed data, and values of $\alpha < 1$ result in greater weight on the observed data than on the notional prior. The value for α can be set using either a subjective belief or a validation sample, as described in Section 2.4.

Some maximum likelihood numerical routines require a binary {0,1} dependent variable, so it is not possible to record the artificial observations by adding fractions (e.g., g_{sj}) to the original dependent variable vector. However, these routines frequently allow a vector of weights

[3] For simplicity of exposition, we illustrate our approach assuming a fixed number of choice tasks S and fixed number of alternatives per choice task J; however, this is not a requirement of the approach.

to be applied to the observations. The prior described in Eq. (3) can be implemented using standard choice modeling software such as STATA, NLOGIT, BIOGEME, and Latent Gold by augmenting the observed data with artificial observations and then weighting the artificial observations with respect to the original observations, which will result in the modified likelihood (Eq. (4)). Then, Eq. (4) can be maximized with respect to parameters β using standard maximum likelihood techniques.

Table 1 shows how the data can be set up to include a flat prior and a binary dependent variable. The table shows a hypothetical data set in stacked format, where each row corresponds to a single alternative in a single choice task. The artificial observations are added to the observed data of S choice tasks each with J alternatives described by K explanatory variables by replicating J times the $SJ \times K$ matrix of explanatory variables $\mathbf{x}$. The dependent variable vector for the artificial observations $\{y_{rj}|r = S+1, \ldots, S+JS, j=1, \ldots, J\}$ is composed such that each alternative (row) from the original explanatory variable matrix is chosen once, as shown in Table 1. The artificial observations should be weighted by a factor of α / J in the log likelihood function relative to the observed data.

The SAS procedure 'mdc' does not allow weighted observations, but the modified maximum likelihood method can be implemented by replicating the observed or artificial data an appropriate number of times to balance the observed and artificial data according to the desired α. For example, for $g_{sj} = 1 / J$, i.e., a flat prior, and values $\alpha \leq 1$, the data should be composed first as in Table 1. Then, the observed data should be replicated a number of times $I = J / \alpha$. For a flat prior and values $\alpha > 1$, the data should be composed first, as in Table 1. Then, the artificial data should be replicated a number of times $I = J / \alpha$. Necessarily, I is restricted to integer values in these cases.

2.2. Alternatives for specifying prior

We now demonstrate two alternative forms of the conjugate prior, corresponding first to a notional sample that uses the same choice tasks as the discrete choice experiment but with a sample proportional prior, or empirical Bayes approach, and second to a notional sample composed of the same choice tasks as the discrete choice experiment and additional choice tasks of market relevant alternatives. The market relevant prior is not operational when new-to-the-market alternatives and attributes are to be tested. In these cases, the market relevant prior approach could be adapted instead using fictitious choice tasks that would be a representation of an investigator's subjective belief about the future market.

Table 1
Stacked data format showing observed data and artificial observations for estimation with binary dependent variables and weighted choice tasks.

	Choice tasks (s)	Alt. (j)	Choice (y)	x_1	x_2	x_3	Log likelihood weight	
Observed data	1	1	0	1	1	1	1	
	1	2	0	−1	−1	1	1	
	1	3	1	−1	−1	−1	1	
	2	1	0	−1	1	−1	1	
	2	2	0	−1	−1	1	1	
	2	3	1	1	1	−1	1	
								From choice task s
Artificial data	3	1	1	1	1	1	α / J	1
	3	2	0	−1	−1	1	α / J	1
	3	3	0	−1	−1	−1	α / J	1
	4	1	0	1	1	1	α / J	1
	4	2	1	−1	−1	1	α / J	1
	4	3	0	−1	−1	−1	α / J	1
	5	1	0	1	1	1	α / J	1
	5	2	0	−1	−1	1	α / J	1
	5	3	1	−1	−1	−1	α / J	1
	6	1	1	−1	1	−1	α / J	2
	6	2	0	−1	−1	1	α / J	2
	6	3	0	1	1	−1	α / J	2
	7	1	0	−1	1	−1	α / J	2
	7	2	1	−1	−1	1	α / J	2
	7	3	0	1	1	−1	α / J	2
	8	1	0	−1	1	−1	α / J	2
	8	2	0	−1	−1	1	α / J	2
	8	3	1	1	1	−1	α / J	2

2.2.1. Empirical Bayes prior

The empirical Bayes approach uses the choice tasks of the discrete choice experiment to form the notional sample, as explained in Section 2.1 for the case of a flat prior. The difference between the flat prior and the empirical Bayes approach is the assignment of the weight proportions g_{sj} on the prior. The empirical Bayes approach first uses the data to calculate the weight proportions for the prior before using the data again to estimate the model (Carlin & Louis, 2000). For example, Clogg et al. (1991) take the prior weight as being proportional to the frequency of each response category outcome observed across their entire sample.

In keeping with this style of empirical Bayes approach, which is equivalent to shrinking to the observed population mean choice probabilities rather than shrinking to equal probabilities, a similar method can be applied to discrete choice models when a sufficient number of individuals N complete a set of identical choice tasks S, as in a discrete choice experiment. The sample proportion

$$\overline{P}_{sj} = 1/N \sum_{i=1}^{N} y_{isj} \tag{6}$$

of chosen alternative j in choice task s can be used to modify the prior weights proportionally: $g_{sj} = \overline{P}_{sj}$, rather than use a constant weight proportion $g_{sj} = 1 / J$ for all alternatives in all choice tasks. For cases when an alternative j in a particular choice task s is never chosen by the sample population, a small value such as $1 / N$ can be applied to the unchosen alternative in place of $\overline{P}_{sj}$.

For the empirical Bayes prior, assign a positive weight α that will apply to each choice task s and assign weight proportions $g_{sj} = \overline{P}_{sj}$. The conjugate prior is given by Eq. (3), the posterior is given by Eq. (4), and R_{prior} is given by Eq. (5). As in the case of the flat prior, values of $\alpha > 1$ result in greater weight on the notional prior than on the observed data, and values of $\alpha < 1$ result in greater weight on the observed data than on the notional prior. The value for α can be set using either subjective beliefs or a validation sample, as described in Section 2.4.

2.2.2. Market relevant prior

For the market relevant notional sample, begin with the flat prior and then append $q = 1, \ldots, Q$ market relevant choice tasks composed of $m = 1, \ldots, M$ alternatives, where each alternative m is described by a $k \times 1$ vector of explanatory variables $\mathbf{x}_{qm}$. Assign weight α to the choice tasks representing the flat prior and the market relevant choice tasks. Assign weight proportions $g_{sj} = 1 / J$ to the choice tasks representing the flat prior and weight proportions g_{qm} to each alternative m in market relevant choice task q according to the observed market shares or subjective belief, such that $\sum_{m=1}^{M} g_{qm} = 1, q = 1, \ldots, Q$.

The conjugate prior is given by

$$p(\beta) \propto \prod_{s=1}^{S} \prod_{j=1}^{J} \pi_{sj}^{\alpha g_{sj}} \prod_{q=1}^{Q} \prod_{m=1}^{M} \pi_{qm}^{(S/Q)\alpha g_{qm}}, \tag{7}$$

where α is a positive constant to be specified and the expected choice shares π_{sj} and π_{qm} are calculated according to Eq. (2). The factor S / Q in the expression for the market relevant choice tasks results in the contribution of the prior to the likelihood function being equally divided between the notional sample representing the flat prior and the notional sample representing the market, regardless of how many market relevant choice tasks are included.

Combining the flat prior with the market relevant choice tasks, rather than using solely the market relevant choice tasks, is a precaution to avoid data separation, which may occur when the number of market relevant choice tasks Q is small. The notional sample for the flat prior and the market relevant choice tasks can be combined to form a conjugate prior in many ways, including specifying different weights α for the flat prior and the market relevant choice tasks. The specification in Eq. (7) is only one example.

Combining the prior with the conventional likelihood function from Eq. (1), the posterior is then

$$p(\beta|y) \propto \prod_{s=1}^{S}\prod_{j=1}^{J}\pi_{sj}^{y_{sj}+\alpha g_{sj}}\prod_{q=1}^{Q}\prod_{m=1}^{M}\pi_{qm}^{(S/Q)\alpha g_{qm}}. \tag{8}$$

Given that there are S original observations, each with a weight y_{sj} of unity, and JS artificial observations, each with a weight of $\alpha g_{sj} = \alpha / J$, and MQ artificial observations, each with a weight of $(S / Q)\alpha g_{qm}$, the ratio R_{prior} of the artificial observations, i.e., the prior as described by a notional sample, to the original data observations is

$$R_{prior} = \frac{\alpha S + (S/Q)\alpha Q}{S} = 2\alpha. \tag{9}$$

Therefore, values of $2\alpha > 1$ result in greater weight on the notional prior than on the observed data, and values of $2\alpha < 1$ result in greater weight on the observed data than on the notional prior. The value for α can be set using either subjective beliefs or a validation sample, as described in Section 2.4.

2.3. Model evaluation

We wish to assess the modified maximum likelihood approach according to parameter recovery for simulated data and predictive accuracy for both simulated and observed data. We compute the root mean squared error $RMSE$ of each parameter β_k and the parameter estimate $\hat{\beta}_k$ as a measure of parameter recovery (see Andrews, Ainslie, and Currim (2002) for a similar application of this measure):

$$RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\hat{\beta}_k - \beta_k\right)^2}, \tag{10}$$

where K variables describe each choice alternative.

We consider two cases for which we wish to assess predictive accuracy. The first case concerns new choice tasks for the sample individuals used in estimation. The second case involves choice tasks answered by individuals other than those in the sample used for estimation.

For new choice tasks for the individuals used in estimation, we compute the root likelihood RLH for a series of T holdout choice tasks as a measure of predictive accuracy:

$$RLH = \left\{\prod_{i=1}^{N}\prod_{t=1}^{T}\prod_{j=1}^{J}\pi_{itj}\left(\hat{\beta}_i\right)^{y_{itj}}\right\}^{\frac{1}{NT}}, \tag{11}$$

where the predicted choice probability $\pi_{itj}\left(\hat{\beta}_i\right)$ is computed as in Eq. (2) for the set of T holdout choice tasks using the individual estimates of the parameters $\hat{\beta}_i$ for each individual i. The RLH prediction measure compares the prediction $\pi_{itj}\left(\hat{\beta}_i\right)$ for respondent i facing alternative j in new choice set t to the observed choice y_{itj}. Root likelihood is a transformation of the likelihood function that normalizes the likelihood value based on the number of choice tasks and can be interpreted as the geometric mean of the model predicted choice probabilities for the observed chosen alternatives.

The likelihood function as applied in Eq. (11) is one example of a scoring function, and many scoring functions can be posited (Gneiting & Raftery, 2007). A scoring function is a measure of model prediction performance that relates the probabilistic model predictions to the events that actually occur.

In contrast to the root likelihood (Eq. (11)) that is computed for predictions of the estimation sample of individuals facing new choice tasks, a measure of model prediction performance can be calculated for the more general case of predicting the choices of a holdout sample of individuals facing new choice tasks. We label our measure as a root predictive score (RPS) to score the model on its predictions of the T choices of H holdout individuals. It is a measure of the accuracy of the predicted choice probabilities for holdout sample H based on the model average predictions for individual h made from the N individual models. The root predictive score is

$$RPS = \left\{\prod_{h=1}^{H}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\prod_{t=1}^{T}\prod_{j=1}^{J}\pi_{itj}\left(\hat{\beta}_i\right)^{y_{htj}}\right)\right]\right\}^{\frac{1}{HT}}, \tag{12}$$

for a population of H holdout individuals facing T new choice tasks each with J alternatives, given an estimation sample of N individuals, where $y_{htj} = 1$ when holdout individual h chooses alternative j in choice task t and $y_{htj} = 0$ otherwise and $\pi_{itj}\left(\hat{\beta}_i\right)$ is the predicted choice probability computed by Eq. (2) for alternative j in choice task t according to model i.

To understand the interpretation of the root predictive score, consider the following. Given a holdout sample of individuals H, we do not know how to weight the N individual models developed from the estimation sample to best predict the sequence of choices of a particular individual h in the holdout sample. Therefore, we apply model averaging and give equal weight to the individual models $n = 1, \ldots, N$. The term inside the parentheses is the predictive score for the T choices of a particular holdout individual h according to model i. The term inside the brackets averages the predictive scores over each of the N models from the estimation sample. The term inside the braces is the predictive score for the entire holdout sample H using the model average predictive scores for each holdout individual h. The quantity within the braces is raised to the power $1 / (HT)$, which results in the root predictive score over holdout individuals H and new choice tasks T. The root predictive score always takes values between zero and one and is a measure of model prediction capability for a holdout sample facing new choice tasks, with $RPS = 1$ indicating perfect prediction.

2.4. Prior weight calibration

The modified maximum likelihood method (MML) described in Section 2.1 requires the specification of the weighting parameter α to balance the notional prior and the observed data. Whether applied to a single individual or to a sample, the best value for α varies from problem to problem, based on the similarity of the prior and the observed data and based on available information, e.g., the number of choice tasks, contained in the observed data.

We consider two approaches for specifying the prior weight α used in model estimation. First, subjective judgment based on experience can provide structure for the problem and should inform the choice of α. For example, a value for α can be chosen based on values of α that have worked well for similar problems.

Second, a numerical routine can be implemented to find the prior weight that maximizes model prediction performance. We would like to calibrate the model to maximize the root predictive score (Eq. (12)) because it coincides with the common marketing objective of using data from one sample to predict the choices of a new sample facing new choice tasks. However, we would also like to formulate an estimation method that does not require us to abandon valuable choice tasks and respondents solely for calibration.

Therefore, the numerical approach we propose for setting the prior weight α has similarities to formal cross-validation methods. The procedure is as follows:

1. Conduct the modified maximum likelihood estimation V times for a particular value of α where for each estimation $v = 1,\ldots,V$ a random subsample of half the original sample is withheld from the estimation.
2. Calculate the root predictive score for the withheld subsample according to Eq. (12), except the new choice tasks T in Eq. (12) are replaced by the choice experiment choice tasks S.
3. Average the resulting values for *RPS* over the V estimations.
4. Repeat the procedure for a finite set of prior weight values α.
5. Select the value for α that results in the highest average *RPS*.

For the flat prior, we recommend testing values for α in the range $0.001 \leq \alpha \leq 4$ and using a greater density of trials for the range $0.05 \leq \alpha \leq 0.5$. Once a prior weight α has been selected, the models can be estimated using the entire data set. Sections 3 and 4 use simulation studies and data from four discrete choice experiments to show the impact of different values of α on the estimation results.

2.5. Method summary

Fig. 1 illustrates the steps required to implement the method. The steps are as follows: collect the data; construct a notional sample; combine the observed data with the notional sample; calibrate the weight for the prior, or notional sample, versus the observed data; and estimate the models with the chosen prior weight value.

3. Method evaluation using Monte Carlo simulations

In this section, we use Monte Carlo simulations to consider the properties of the modified maximum likelihood method in a situation typical of the analysis of discrete choice experiments. The Monte Carlo simulations consider two factors: the weight placed on the prior and the number of choice task observations used in estimation.[4]

The performance metrics are the measure of parameter recovery *RMSE* and either of the two measures of the prediction of new choice tasks *RLH* or *RPS*. The *RLH* measure emphasizes the prediction of new choices faced by the same individuals in the estimation sample, while the *RPS* measure emphasizes the performance of the estimation sample models in predicting to a holdout sample of individuals in either new choice tasks or the same choice tasks faced by the estimation sample. The simulations in Section 3.2 consider the performance of the method for single individuals, and therefore report *RLH*. The simulations of Section 3.3 and the results of Section 4 consider the out-of-sample performance of the estimated models and therefore report *RPS*.

The Monte Carlo experiment simulates an indirect utility function with seven attributes. Attributes $z = 1$ and $z = 2$, which could represent brand and price, each assume one of four levels ($L_z = 4$). Attributes $z = 3, 4, 5, 6, 7$, representing other features, each assume one of two levels ($L_z = 2$). For model estimation, we use effects coding of choice alternative $\mathbf{x}$. An alternative $\mathbf{x}$ is defined as a vector where the length of the vector is $K = \sum_{z=1}^{Z}(L_z - 1) = 11$. We use indices $z = 1, \ldots, Z$ and $\ell = 0,\ldots,L{-}1$ to define the effects coding. If attribute z assumes level $\ell = 0$, then $x_{z\ell'} = -1 \forall \ell' = 1,\ldots,L_z{-}1$. If attribute z assumes level $\ell \neq 0$, then $x_{z\ell} = 1$ and $x_{z\ell'} = 0 \forall \ell' \neq \ell, \ell' \in \{1,\ldots,L_z{-}1\}$. Utility is linear in attributes and an error term ε:

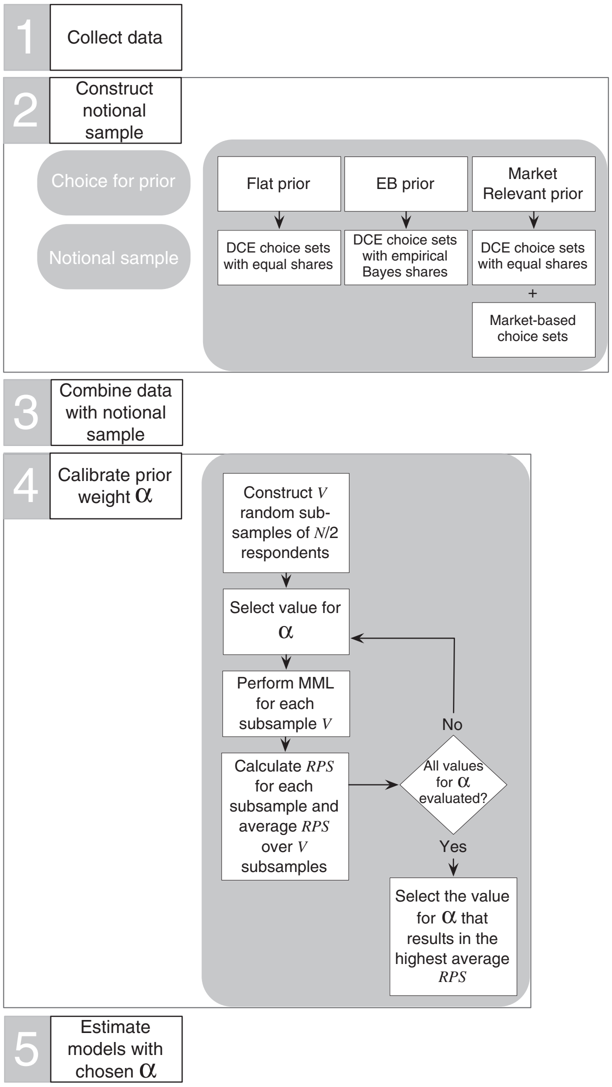
1
Collect data
2
Construct notional sample
Choice for prior
Notional sample
Flat prior
EB prior
Market Relevant prior
DCE choice sets with equal shares
DCE choice sets with empirical Bayes shares
DCE choice sets with equal shares
+
Market-based choice sets
3
Combine data with notional sample
4
Calibrate prior weight α
Construct V random sub-samples of N/2 respondents
Select value for α
Perform MML for each subsample V
Calculate RPS for each subsample and average RPS over V subsamples
All values for α evaluated?
No
Yes
Select the value for α that results in the highest average RPS
5
Estimate models with chosen α

Fig. 1. Steps for implementing modified maximum likelihood method with conjugate priors.

$$U = \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \beta_4 x_{21} + \beta_5 x_{22} + \beta_6 x_{23} + \beta_7 x_{31} + \beta_8 x_{41} + \beta_9 x_{51} + \beta_{10} x_{61} + \beta_{11} x_{71} + \varepsilon. \tag{13}$$

The objective of the modified maximum likelihood approach is to find parameter estimates for β given data from a discrete choice experiment. In our simulated experiment, each choice task is composed of four choice alternatives. The respondent's utility for each alternative is Eq. (13), with the random variable ε independently and identically distributed as extreme value type I across all choice alternatives and tasks. Denoting the 11×1 vector of parameter estimates by β and the vector of covariates for choice alternative j of choice task s by $\mathbf{x}sj$, the probability π_{sj} is given by Eq. (2).

The population distributions from which the coefficient vector β is drawn are normal. In every case, the coefficients are independently

[4] The distribution from which the simulated data is generated was another simulation factor. However, the results from four other distributions are omitted for brevity because they followed the same patterns as the reported results.

Table 2
Simulated parameter normal distribution specification used in the Monte Carlo simulation studies.

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}
μ	0	−0.4	−0.5	1.9	−0.6	−2.7	0.6	0.5	1.4	1.9	−1.2
σ	0.6	0.66	0.6	0.3	0.45	1.05	0.36	0.3	0.48	0.51	0.39

but not identically distributed. Table 2 provides the specification of the distribution for each coefficient.[5]

The Monte Carlo experiment simulates the distribution of the modified maximum likelihood estimate of β based on Eqs. (13) and (2) and on the choices of single respondents. The simulated data in the Monte Carlo experiment are generated for five choice experiment designs that correspond to different numbers of choice tasks. The numbers of choice tasks S in which choices are made among four alternatives are $S \in \{4,8,16,32,64\}$. For each choice experiment, there are 1000 simulated individuals, and modified maximum likelihood estimates are constructed for each simulated individual.

For each simulated choice experiment, estimation is undertaken for several values of α. The range of values for α is chosen to extend from a near-zero weight prior $\alpha = 0.001$ (i.e., the notional prior data contribution is 0.1% of the observed data contribution) to a heavily weighted prior $\alpha = 4$ (i.e., the notional prior data contribution is four times the observed data contribution). Intuitively, less information is needed from the prior for the choice experiments with more choice tasks. Therefore, the values of α used for the different simulation conditions differ with different numbers of choice tasks to concentrate the simulation results in the region where the parameter recovery and holdout prediction measures attain their best values for this particular problem.

We choose choice experiment designs that are orthogonal or nearly orthogonal in main effects for the Monte Carlo simulations. This approach corresponds to the (unrealistic) assumption that the vector of part-worths is zero (Kessels, Jones, Goos, & Vandebroek, 2011a,b). We make this choice because it represents a common approach used by practitioners (Chrzan & Orme, 2000) and in many academic disciplines, including transport (Bliemer & Rose, 2011) and health economics (de Bekker-Grob, Ryan, & Gerard, 2012). Although Bayesian efficient designs offer superior statistical efficiency (Sándor & Wedel, 2001, 2005), there are barriers to their adoption, including tradition, a lack of widely distributed software,[6] and a lack of Bayesian efficient design expertise. While we recommend statistically efficient designs when available, it is important to note that the modified maximum likelihood estimation methodology described here does not depend on a specific choice experiment design methodology. The estimation method can be applied to data from any discrete choice experiment with sufficient observations per respondent, including randomly generated designs. It also appears to be a straightforward extension to adapt the estimation method to non-experimentally generated data, such as from scanner panels; this extension is left for future study. Section 5 discusses the implications for the estimation methodology under different design assumptions, including Bayesian efficient designs.

The five designs used in the Monte Carlo simulations were each formed by first taking an orthogonal array of profiles from the SAS catalog (i.e., sixteen profiles for the four choice task design, thirty-two profiles for the eight and sixteen choice task designs, and 128 profiles for the thirty-two and sixty-four choice task designs). The orthogonal arrays consist of rows of product attribute profiles, with each column corresponding to an attribute and each entry corresponding to the level of the particular attribute. Next, we used the %choiceeff macro in SAS, which implements a modified Fedorov candidate-set-search algorithm as the search procedure for finding improved candidate designs according to the D-error criterion (Kuhfeld & Tobias, 2005), assuming the vector of part-worths is zero, to assign profiles to choice tasks of four alternatives each. The D-error is a widely used criterion in the design of choice experiments (Kessels, Goos, & Vandebroek, 2006).

The role of simulation is to test the modified maximum likelihood estimation method in challenging circumstances that would be difficult or impossible to replicate in a discrete choice experiment or in the market. For example, it is impractical to collect large numbers of holdout observations on which to test the predictive capability of the model. However, this task is easily accomplished with simulation.

We choose to evaluate as many holdout tasks as possible to observe model performance over a wide range of circumstances. There are $4^2 \cdot 2^5 = 512$ possible product profile combinations, given the number of attributes and levels for our simulation. Removing the profiles that were used in any of the choice experiment designs, there were sufficient unused profiles to generate 89 holdout choice tasks of four alternatives each. The holdout choice tasks were constructed by random sampling without replacement from the remaining profiles. The holdout tasks were used to compare consistently the out-of-sample prediction performance across the different choice experiments and different prior weights α.

3.1. Maximum likelihood estimation

Before performing Monte Carlo simulations using the modified maximum likelihood technique, estimation with traditional maximum likelihood sheds light on the extent of the data separation problem in small samples for problems typical of discrete choice modeling applications. Lesaffre and Albert (1989) prove that data separation results in the divergence of the ratio of the standard errors of the parameter estimates between a later iteration and an initial step of maximum likelihood estimation. This ratio can be used as a test for data separation with large ratios indicating data separation.[7]

We performed the maximum likelihood estimation for 1000 simulated individuals for each of the five Monte Carlo experimental conditions, which corresponded to the different numbers of choice tasks. For each experimental condition, Table 3 lists the portion of simulated individuals that returned a ratio of the final standard deviation[8] to initial standard deviation greater than 500 for at least one parameter estimate. The percentage of large ratios is at or near 100% for the $S = 4$ and $S = 8$ choice task cases. In these cases, we expect nearly all simulated individuals to exhibit complete data separation or near complete data separation. The percentage of individuals with data separation decreases from sixteen to sixty-four choice tasks, but approximately 3% of

[5] We chose the distribution for parameters $\beta_4 - \beta_6$ to correspond roughly to a four-level price attribute. There is a reversal of preference between the lowest and second-lowest price attribute ($\mu = 1.4$ vs. 1.9), but the increasing levels of price are otherwise monotonically ordered (i.e., μ : 1.4, 1.9,−.6, − 2.7.). These parameters also have moderate standard deviations. The result is that, even with independent drawings from each parameter, most individuals exhibit a relatively flat preference for price between the first two price levels and then monotonically decreasing preferences for price for increasing price levels. This is a conceivable configuration of preferences, especially if there is a latent dimension of quality associated with price that is not captured by the other attributes—a common occurrence in many product categories and many DCEs. That there are some simulated individuals that violate economic theory, if these parameters were strictly interpreted as the price parameters, is not of consequence to the demonstration of the methodology, and it corresponds to the empirical observation of non-monotonically ordered individual estimates of the price parameters in many choice experiments when the price parameters are estimated using effects codes.

[6] The JMP software (a product of SAS) and the Ngene software (a stand-alone package) are both capable of generating Bayesian efficient designs. The license fees for these software compare favorably with other statistics or econometrics packages.

[7] An external check is necessary because common maximum likelihood routines do not identify the estimation problem as being unbounded because one of the numerical convergence criteria may be satisfied, such as minimum change in the objective function. This issue was also noted by Beggs et al. (1981) and Albert and Anderson (1984) for logistic regression and by Lesaffre and Albert (1989) for logistic discrimination.

[8] The final iteration is defined as the one where the maximum likelihood estimation terminated under standard termination conditions. The numerical termination conditions will likely affect the standard error ratios. However, our simulations of our problem show that all simulated individuals with ratios greater than five had ratios greater than 500, illustrating divergence.

Table 3
Percentage of simulated individuals from Monte Carlo simulations with large ratios of final to initial standard errors, indicating data separation.

Number of choice tasks	4	8	16	32	64
Percentage with data separation	100.0%	99.5%	91.2%	45.0%	3.0%

simulated individuals still exhibit data separation for the case of sixty-four choice tasks.

3.2. Modified maximum likelihood estimation

We illustrate the performance of the modified maximum likelihood method with each of the three priors discussed in Section 2. The simulations illustrate that the optimal prior weight is problem dependent. The initial simulations using the flat prior illustrate how the optimal prior weight ratio R_{prior} depends on the number of choice tasks. The subsequent sections illustrate how the optimal prior weight changes with the choice of prior. The online discrete choice experiment examples in Section 4 show that the optimal prior weight depends not only on the number of choice tasks but also on the number of parameters and the distribution of those parameters. The optimal value of R_{prior} provides an interesting insight into the value of the collected data. A smaller optimal R_{prior} indicates a greater relative value of the collected data.

Section 3.3 evaluates the performance of the estimation method when the prior weight calibration method presented in Section 2.4 is used to select the prior weight.

3.2.1. Flat prior

First, we describe the modified maximum likelihood results for the Monte Carlo simulations assuming a flat prior as detailed in Section 2.1.

For 1000 simulated individuals in each experimental condition, we computed the parameter estimates $\hat{\beta}$, the measure of parameter recovery *RMSE* (smaller is better), and the measure of prediction performance *RLH* (larger is better). Fig. 2 reports the average values of *RMSE* and *RLH* for each particular number of choice tasks S and particular prior weight α. Fig. 2(a) shows the average root mean squared error *RMSE* between estimate $\hat{\beta}$ and simulated β parameters over 1000 simulated individuals for each Monte Carlo experimental condition for $0.001 \leq R_{prior} \leq 4$. A lower value on the y-axis indicates that the parameter estimates are closer to the simulated parameters.

The influence of the prior on the parameter estimates increases with increasing R_{prior} from left to right. Values of $R_{prior} = 1$ indicate that half of the data used in the estimation came from the artificial observations (the notional sample), supporting the prior that all outcomes are equally likely. The best weight ratio R_{prior} for minimizing root mean squared error increases slightly with the decreasing number of choice tasks, as we would expect.

Fig. 2(b) shows the average root likelihood *RLH* for the holdout choice tasks over 1000 simulated individuals for each Monte Carlo experimental condition for $0.001 \leq R_{prior} \leq 4$. The root likelihood is a measure of how the estimated model is performing in predicting choice outcomes in new choice tasks, so larger values for *RLH* are preferred. The influence of the prior increases from left to right. The best weight ratio R_{prior} for maximizing the average root likelihood increases noticeably with a decreasing number of choice tasks.

Recalling that it is not possible to evaluate the simulation results at $R_{prior} = 0$ for cases of data separation, the plot shows the behavior of the performance metrics as R_{prior} approaches zero. In all cases, the performance metrics are worse for the lowest values of R_{prior} tested than for some other larger value of R_{prior}. This finding means that in addition to making individual estimation possible for those respondents that exhibit data separation, a positive value of R_{prior} has added value because it produces a lower *RMSE* and higher *RLH*.

3.2.2. Empirical Bayes prior

We describe the modified maximum likelihood results for the Monte Carlo simulations assuming the empirical Bayes prior $g_{sj} = \overline{P}_{sj}$, as detailed in Section 2.2. For 1000 simulated individuals in each experimental condition, we computed the parameter estimates $\hat{\beta}$, the measure of parameter recovery *RMSE*, and the measure of prediction performance *RLH*. We treated the choices of the 1000 simulated individuals as the sample for the purposes of calculating $\overline{P}_{sj}$. We reported the average values of *RMSE* and *RLH* for each particular number of choice tasks S and particular prior α. Fig. 3(a) shows the average root mean squared error *RMSE* between estimate $\hat{\beta}$ and the simulated β parameters over 1000 simulated individuals for each Monte Carlo experimental condition for $0.001 \leq R_{prior} \leq 4$. The best weight ratio R_{prior} for minimizing root mean squared error increases noticeably with the decreasing number of choice tasks.

Fig. 3(b) shows the average root likelihood *RLH* for the holdout choice tasks over 1000 simulated individuals for each Monte Carlo experimental condition for $0.001 \leq R_{prior} \leq 4$. The best weight ratio R_{prior} for maximizing the average root likelihood increases noticeably with the decreasing number of choice tasks.

Similar to the flat prior results, there are a range of positive prior weights R_{prior} that improve the performance measures *RMSE* and *RLH* compared to the lowest values of R_{prior} tested for all five simulated choice experiments.

Applying the empirical Bayes prior, the average *RMSE* (Fig. 3(a)) and average *RLH* (Fig. 3(b)) are smaller and larger, respectively, and the results are less sensitive to the choice of prior weight α compared to the Monte Carlo experiment results that employ the flat prior (Fig. 2).[9] The improvements in both measures from using the empirical Bayes prior compared to the flat prior diminish as the number of choice tasks per respondent increases such that for $S = 64$ choice tasks, the best average *RMSE* and *RLH* are similar for the two different prior distributions. This result is intuitive because, as more information is available, i.e., more observed choice tasks, the performance of the model should depend less on the contribution of the prior and more on the contribution of the observed data. This intuition is supported by the observation that the ratio of artificial data to observed data R_{prior} decreases with increasing choice tasks for both priors.

3.2.3. Market relevant prior

We describe the modified maximum likelihood results for the Monte Carlo simulations assuming a market relevant prior, as detailed in Section 2.2, where an artificial choice task is composed of alternatives available in the market and the alternatives are assigned choice probabilities according to current market shares. We use a single market relevant choice task as an example. There is no limit to the number of choice tasks that can be included in the notional sample for the prior.

For the Monte Carlo experiments, we take our market relevant choice task to be one of the $T = 89$ previously defined holdout tasks, and we take the choice probabilities to be the observed choice probabilities from a simulation of 100,000 individuals. Table 4 lists the explanatory variables that define the choice task and the corresponding choice probabilities.

For 1000 simulated individuals in each experimental condition, we computed the parameter estimates $\hat{\beta}$, the measure of parameter recovery *RMSE*, and the measure of prediction performance *RLH*. We reported the average values of *RMSE* and *RLH* for each particular number of choice tasks S and particular prior weight α. Fig. 4(a) shows the average root mean squared error *RMSE* between estimate $\hat{\beta}$ and simulated β parameters over 1000 simulated individuals for each Monte Carlo experimental condition for $0.002 \leq R_{prior} \leq 8$. Recall that in the market relevant specification $R_{prior} = 2\alpha$; Fig. 4(b) thus shows the average root likelihood *RLH*

[9] We define decreased sensitivity for a particular choice experiment as an increase in the magnitude of the range of prior weights R_{prior} for which the value of *RLH* is within 5% of the peak value.

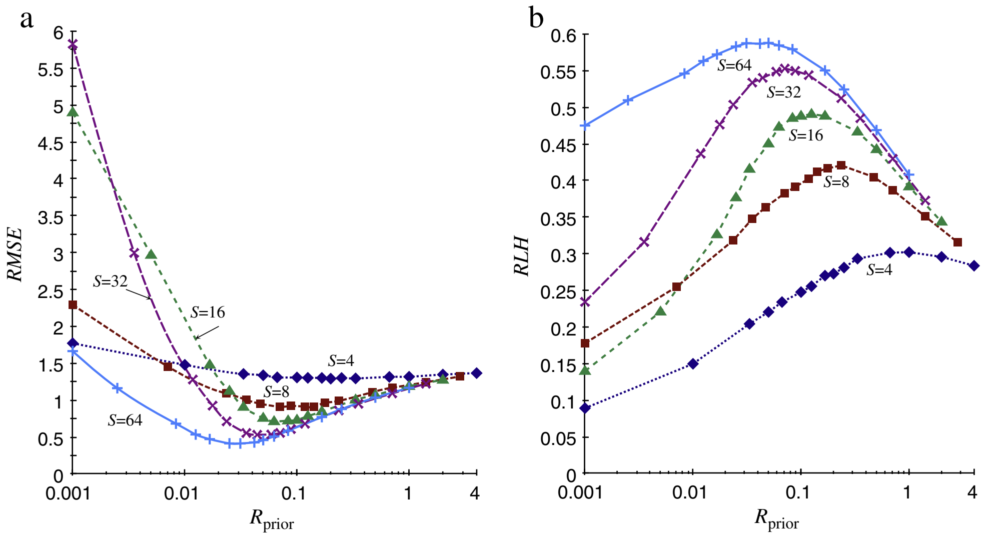

Fig. 2. (a) The root mean squared error *RMSE* between the parameter estimates and the simulated parameters for the flat prior averaged over 1000 simulated individuals for each experimental condition as a function of the ratio of artificial data to observed data R_{prior} used in estimation. (b) The root likelihood *RLH* for the $T = 89$ holdout choice tasks for the flat prior averaged over 1000 simulated individuals for each experimental condition as a function of the ratio of artificial data to observed data R_{prior} used in estimation.

for the holdout choice tasks over 1000 simulated individuals for each Monte Carlo experimental condition for $0.002 \leq qR_{prior} \leq 8$.

The best weight ratio R_{prior} for minimizing *RMSE* increases noticeably with the decreasing number of choice tasks, and the best weight ratio R_{prior} for maximizing the average *RLH* increases noticeably with the decreasing number of choice tasks. Similar to the previous two priors, there are positive values for R_{prior} that result in improved performance measures *RMSE* and *RLH* compared to the lowest values of R_{prior} tested for all five simulated choice experiments.

Applying the market relevant prior, the average root likelihood (Fig. 4(b)) is larger and the result less sensitive to the choice of prior weight compared to the Monte Carlo experiment results that employ the flat prior (Fig. 2(b)). The performance of the market relevant prior, in terms of the lowest *RMSE* and highest *RLH*, is between the performance of the flat prior and the empirical Bayes prior for $S \in \{4,8,16\}$ choice tasks. This result is intuitive because going from the flat prior to the market relevant prior to the empirical Bayes prior increases the information available regarding the sample mean behavior, thus allowing greater shrinkage toward the mean in estimation. Similar to the empirical Bayes prior, the improvement in root likelihood using the market relevant prior compared to the flat prior diminishes as the number of choice tasks per respondent increases, so that for $S \in \{32,64\}$ choice tasks, the best average root likelihood on the holdout choice tasks achieved is the same as that achieved for the flat prior and the market relevant prior distributions.

Careful observation of Figs. 2–4 shows that the prior weights that optimize *RMSE* are, in general, different from those that optimize *RLH*. This is not unexpected because *RMSE* is a symmetric measure that does not differentiate between parameter values that are larger or smaller than the true value and does not differentiate between error in a parameter that has a larger influence on the choice prediction and one that does not. The *RLH* measure is the geometric mean of the predicted probabilities for the chosen alternatives and is penalized asymmetrically for overly optimistic predictions. The result is that

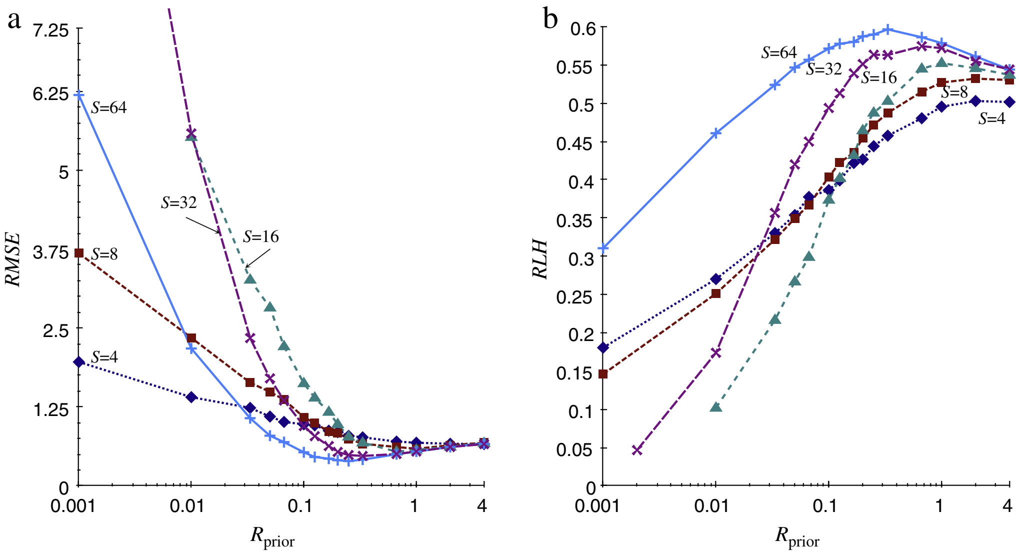

Fig. 3. (a) The root mean squared error *RMSE* between the parameter estimates and the simulated parameters for the empirical Bayes prior averaged over 1000 simulated individuals for each experimental condition as a function of the ratio of artificial data to observed data R_{prior} used in estimation. (b) The root likelihood *RLH* for the $T = 89$ holdout choice tasks for the empirical Bayes prior averaged over 1000 simulated individuals for each experimental condition as a function of the ratio of artificial data to observed data R_{prior} used in estimation.

Table 4
Artificial market relevant choice task.

Alternative (J)	Choice share g_{qm}	x_{11}	x_{12}	x_{13}	x_{21}	x_{22}	x_{23}	x_{31}	x_{41}	x_{51}	x_{61}	x_{71}
1	0.663	0	0	1	−1	−1	−1	1	1	1	1	1
2	0.052	0	0	1	0	1	0	−1	1	1	1	1
3	0.064	0	0	1	0	1	0	1	−1	−1	1	−1
4	0.221	−1	−1	−1	1	0	0	1	1	1	−1	1

Table 5
The optimal prior weight for *RMSE* and *RLH* from the Monte Carlo simulations for each of the three priors tested.

Choice tasks	Flat		Empirical Bayes		Market relevant	
	α_{RMSE}	α_{RLH}	α_{RMSE}	α_{RLH}	α_{RMSE}	α_{RLH}
4	0.2	1	4	4	0.4	1.33
8	0.141	0.236	1	2	0.2	0.5
16	0.0625	0.125	0.666	1	0.133	0.25
32	0.0442	0.0707	0.333	0.667	0.1	0.133
64	0.0313	0.05	0.25	0.333	0.0667	0.0667

predictive performance is improved when prior weights are slightly larger than the prior weight that nominally improves parameter fit. As expected, the discrepancy in the optimal prior weight for both measures diminishes as the number of observations increases. Table 5 lists the optimal prior weight for each performance measure for each of the three priors tested.

3.3. Specification of prior weight

The data generating process is unknown and large numbers of holdout tasks are rarely available in application. Instead, the choice of prior weight α must be made using only the data at hand. The calibration method described in Section 2.4 is designed for this situation. We can use Monte Carlo simulation studies and online choice experiment data with true holdout individuals and choice tasks to provide an indication of how well the calibration procedure will work in practice. This section, therefore, illustrates a test of the prior weight calibration method described in Section 2.4.

In this section, we refer to two different sets of choice tasks: design tasks and holdout tasks. Design tasks are used in estimation and in the calibration procedure. Holdout tasks are additional choice tasks that are completed by each individual, whether simulated or real. These tasks are not used in estimation or validation. Instead, they are used to assess predictive validity as a surrogate for the market where we would like to predict purchases, for example.

To test the calibration procedure, we first created a number of estimation and validation subsample pairs, as described in Section 2.4. Next, for each subsample pair, we found the recommended prior weight α_{rec} given the estimation sample and design choice tasks using the root predictive score of the validation subsample over the design tasks. Then, we computed the predictive performance $\left(RPS|_{\alpha_{rec}}\right)$ for the validation subsample over the holdout choice tasks. This is in contrast to the root predictive score computed for the validation subsample over design choice tasks that is used in the calibration procedure to calibrate the prior weight value. We also found the optimal prior weight value α_{best} that maximizes the predictive performance $\left(RPS|_{\alpha_{best}}\right)$ of the model estimates for the validation subsample over the holdout tasks. This is equivalent to performing the calibration procedure using the root predictive score from the validation subsample over the holdout choice tasks rather than over the design choice tasks. We did this procedure to have a measure for model predictive validity performance for a sample that was not used in estimation and for choice tasks that were not used in estimation. Finally, we compared the predictive performance on the validation sample over the holdout tasks between the model estimates using α_{rec} and the model estimates using α_{best}. We repeated this procedure for each subsample pair and then averaged the results. Small differences between the two predictive performance values ($RPS|_{\alpha_{rec}}$ and $RPS|_{\alpha_{best}}$) indicate an accurate calibration procedure.

We conducted the test as follows for each of the experimental designs previously described, i.e., $S \in \{4,8,16,32,64\}$. First, we simulated a sample of 200 individuals by taking random drawings from the parameter distributions. Second, we composed the notional sample assuming a flat prior. Third, we followed the calibration procedure outlined in steps 1–5 from Section 2.4, where we set the number of validation subsamples $V = 10$.

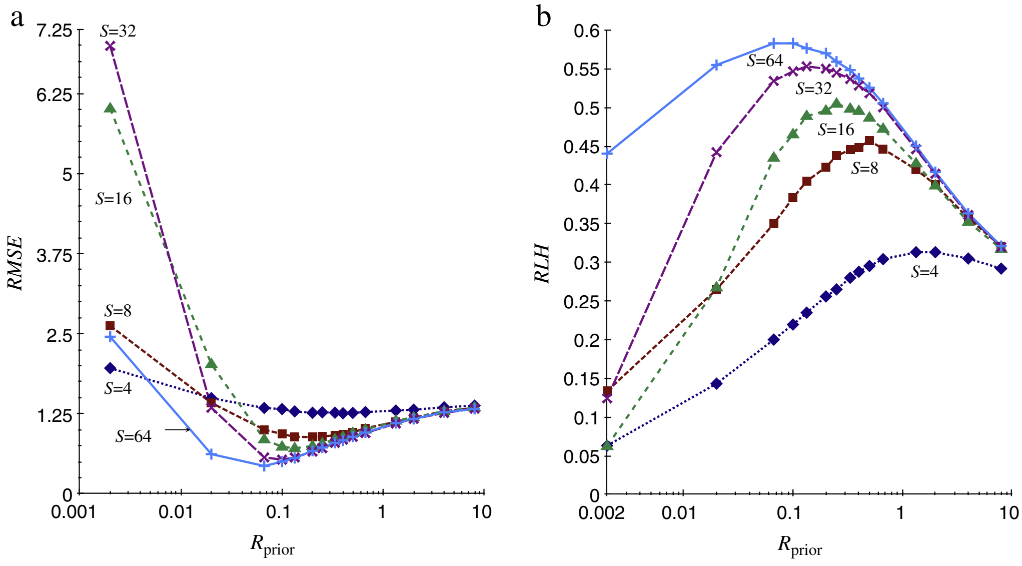

Fig. 4. (a) The root mean squared error *RMSE* between the parameter estimates and the simulated parameters for the market relevant prior averaged over 1000 simulated individuals for each experimental condition as a function of the ratio of artificial data to observed data R_{prior} used in estimation. (b) The root likelihood *RLH* for the $T = 89$ holdout choice tasks for the market relevant prior averaged over 1000 simulated individuals for each experimental condition as a function of the ratio of artificial data to observed data R_{prior} used in estimation.

Table 6
The lowest, best, and recommended prior weights α for each choice experiment, and the average root predictive score $\overline{RPS}$ for each α. The $\overline{RPS}$ values were calculated for the validation samples on the $T = 89$ holdout choice tasks. The final column gives the percentage differences in $\overline{RPS}$ evaluated at the best value of α and the value of α recommend by the prior weight calibration procedure.

Num. choice tasks	α_{low}	α_{best}	α_{rec}	$\overline{RPS}\|_{\alpha_{low}}$	$\overline{RPS}\|_{\alpha_{best}}$	$\overline{RPS}\|_{\alpha_{rec}}$	% diff. in $\overline{RPS}$ @ α_{rec}, α_{best}
S							$100\frac{\overline{RPS}\|_{\alpha_{rec}} - \overline{RPS}\|_{\alpha_{best}}}{\overline{RPS}\|_{\alpha_{best}}}$
4	0.0007	0.67	0.05	0.13	0.32	0.26	−17.7
8	0.0007	0.08	0.10	0.29	0.48	0.47	−0.7
16	0.0007	0.04	0.05	0.32	0.52	0.52	−0.4
32	0.0007	0.06	0.04	0.46	0.55	0.54	−0.9
64	0.001	0.05	0.03	0.52	0.56	0.55	−0.5

The calibration procedure identifies a recommended prior weight α_{rec} that maximizes the root predictive score calculated for the validation subsamples for the S choice tasks corresponding to the choice experiment. Because it is a simulation, we can also calculate the root predictive score for the validation subsample for the $T = 89$ holdout tasks. We perform this calculation to assess the accuracy of the calibration method for use in out-of-sample prediction, i.e., new people and new choices. The calibration method itself uses new people, i.e., the validation sample, but uses the same choice tasks as the estimation sample.

Table 6 shows the lowest prior weight tested α_{low} for a particular design, the prior weight α_{best} that provides the best average root predictive score for the $T = 89$ holdout choice tasks, and the prior weight α_{rec} recommended by the calibration procedure. The next three columns list the average root predictive score $\overline{RPS}$ for the $T = 89$ holdout choice tasks evaluated at the respective α. The final column lists the percentage difference between $RPS|_{\alpha_{rec}}$ and $RPS|_{\alpha_{best}}$. The percentage difference is one measure of how close the prior weight calibration procedure came to identifying a prior weight value that would have yielded the best root predictive score on the holdout tasks.

The prior weight calibration procedure achieves average root predictive scores within 1% of the best average root predictive scores observed, with the exception of the four choice task case where the difference was more than 17%. The larger discrepancy for the four choice task case indicates that one should take care when selecting the prior weight for cases where the degrees of freedom in the data $(S(J-1))$ are near the degrees of freedom of the choice model (K). Table 6 also shows that the average root predictive scores with a very low prior weight α_{low} are much lower than the average root predictive scores observed when more weight is placed on the prior, including the values for α recommended by the calibration procedure.

4. Discrete choice experiment examples

We tested the modified maximum likelihood approach using four data sets from online discrete choice experiments in the U.S. and Australia. The car insurance and airline data sets each had a sample of 200 respondents and twelve choice tasks with four alternatives each for estimation and four holdout choice tasks, and they had the same parameterization as the simulation studies. The four holdout choice tasks had four alternatives each. The pizza data set had a sample of 600 respondents, twenty choice tasks with five alternatives each including a "none" alternative for estimation, and five holdout choice tasks, and it had fifteen parameters corresponding to four four-level attributes, two two-level attributes, and a "none" alternative constant. The digital camera data set had a sample of 600 respondents, twenty-four choice tasks with five alternatives each including a "none" alternative for estimation, and five holdout choice tasks, and it had twenty-five parameters corresponding to one six-level attribute, one five-level attribute, three four-level attributes, two three-level attributes, two two-level attributes, and a "none" alternative constant. The five holdout tasks included a "none" alternative, and four of the holdout tasks included four additional alternatives while the fifth holdout task included eight additional alternatives. Table 7 summarizes the discrete choice experiments, including the percentage of respondents that appear to have separated data as indicated by large standard deviations for the parameter estimates when the termination criterion is reached in conventional maximum likelihood estimation.

4.1. Model estimation

Each data set was divided randomly into an estimation sample and a validation sample of equal size. This procedure was repeated ten times to create ten estimation and validation sample pairs. We followed the procedure described in Section 2.4 to identify a suitable prior weight, first using the flat prior and then using the empirical Bayes prior. The online discrete choice experiment results exhibit the same trends with respect to changes in α as the simulation results. Similar to Table 6, we report in Table 8 the prior weight α_{rec} recommended by the weight calibration procedure and the resulting average root predictive score $\overline{RPS}|_{\alpha_{rec}}$ as well as the best weight α_{best} and best average root predictive score $\overline{RPS}|_{\alpha_{best}}$ for each data set, where the root predictive scores reported are evaluated for the $T = 4$ or $T = 5$ holdout choice tasks. The percentage differences between average root predictive scores evaluated at the recommended and the best prior weights are between 0 and 2%. The best value for the prior weight according to the average root predictive score differs across the data sets and is in the range of 0.02–0.67, while the recommended prior weights are between 0.12 and 0.50.

The empirical Bayes prior has larger prior weights and achieves a small improvement in RPS relative to the flat prior, as we would expect from the simulation study results. The prior weight calibration procedure using either prior achieves average root predictive scores within 2% of the best average root predictive scores observed. These small differences relative to the best prior weight confirm the simulation results and indicate that the cross-validation procedure is working as designed.

4.2. Computation time

Computation time for the method will be proportional to the number of modified maximum likelihood estimations conducted. The computation time is thus proportional to the sample size, the number of sample/validation subsample pairs, and the number of values of α tested. The number of modified maximum likelihood estimations used to collect the data listed in Table 8 is the product of the number of individuals in the estimation subsample (i.e., half the number listed in Table 7), the number of subsample data sets (i.e., ten), and the number of values of α tested (i.e., fifteen). The other significant computation is the calculation of the root predictive score, which is calculated for each validation subsample and each value of α tested. We used the

Table 7
Description of discrete choice experiments, including the percentage of respondents with data separation.

Data set	Sample size N	Choice tasks/resp. S	Alts./choice task J	HO choice tasks/resp. T	Number of params. K	% data sep.
Pizza	600	20	5	5	15	97.0
Camera	600	24	5	5	25	96.3
Car insurance	200	12	4	4	11	91.5
Airline	200	12	4	4	11	92.0

Table 8
The recommended and best prior weight α and average root predictive score $\overline{RPS}$ for the validation samples on the $T = 4$ or $T = 5$ holdout choice tasks for both a flat prior and an empirical Bayes prior. The final column of results for each prior gives the percentage differences in $\overline{RPS}$ evaluated at the best value of α and the value of α recommend by the prior weight calibration procedure.

Flat prior					
Data set	α_{rec}	α_{best}	$\overline{RPS}\|_{\alpha_{rec}}$	$\overline{RPS}\|_{\alpha_{best}}$	Percent difference in $\overline{RPS}$ evaluated at α_{rec} and α_{best}
Pizza	0.17	0.07	0.278	0.284	−2.04%
Camera	0.30	0.15	0.241	0.243	−0.77%
Car insurance	0.19	0.060	0.380	0.384	−0.99%
Airline	0.12	0.02	0.380	0.386	−1.58%
Empirical Bayes prior					
Pizza	0.17	0.10	0.282	0.286	−1.52%
Camera	0.33	0.25	0.242	0.244	−0.63%
Car insurance	0.50	0.67	0.394	0.400	−1.69%
Airline	0.25	0.03	0.395	0.395	−0.03%

estimation code implemented in Matlab 2011b on an Intel Core 2 Quad 2.67 GHz processor with 8 GB of RAM running the Windows 7 operating system. Three hundred maximum likelihood estimations of the pizza data took 25 s. A single calculation of the root predictive score took 0.62 s. The total computation time of the modified maximum likelihood estimation of the pizza data, including the prior weight calibration, took approximately 64 min. Given the relative insensitivity of root predictive score with respect to α near the best value of $\overline{RPS}$, it is plausible to consider only a single estimation/validation subsample pair and fewer test values for α. This would lead to a reduction in computation time of two orders of magnitude below what was achieved for the results in Table 8. The computation time can be further reduced by estimating the individual models in parallel rather than in series and by optimizing the estimation code.

4.3. Comparison to hierarchical Bayes estimation

We can compare the parameter estimates and the prediction performance of the modified maximum likelihood method to conventional estimation techniques, such as the hierarchical Bayes (HB) estimation of a random coefficients logit model and the maximum likelihood estimated homogeneous multinomial logit (MNL). While a formal comparison is left for future study, Fig. 5 plots the modified maximum likelihood estimates for all parameters, assuming the empirical Bayes prior for the full samples based on the recommended prior weight versus the individual means of the HB posterior estimates for the data sets listed in Table 7. The HB estimation assumes a prior of normally distributed parameters and estimates the full variance–covariance matrix among the parameters.

Table 9
The average root predictive scores $\overline{RPS}$ and corresponding standard deviations for 100 validation subsamples for the four discrete choice experiments for the homogeneous MNL (MNL), the HB estimated random coefficients logit (HB), and the modified maximum likelihood individual level MNL with the empirical Bayes prior (MML).

Data set	MNL		HB		MML	
	$\overline{RPS}$	St. Dev.	$\overline{RPS}$	St. Dev.	$\overline{RPS}\|_{\alpha_{rec}}$	St. Dev.
Pizza	0.210	0.00577	0.296	0.0112	0.283	0.00795
Camera	0.210	0.00520	0.248	0.00970	0.244	0.00697
Car insurance	0.390	0.0217	0.404	0.0314	0.403	0.0239
Airline	0.380	0.0253	0.400	0.0263	0.399	0.0276

We observe that the parameter estimates are highly proportional, as measured by R^2 values between 0.73 and 0.91. Fig. 5(b) has the lowest R^2 value and exhibits the most nonlinear relationship between the estimates. The nonlinearity is due exclusively to the parameter estimates representing the outside good or "none" option relative to the product profiles, as is highlighted in the figure. The pizza data set also includes a "none" alternative in each choice task, but the parameter estimates for the outside good parameter do not exhibit the same relationship as the camera data. The larger magnitude estimates of the "none" parameter for the HB estimation may be capturing those individuals who dogmatically, or nearly so, chose the "none" alternative or never chose the "none" alternative. Both the HB and the MML models yield similar *RPS* scores for the camera data, indicating that the difference in estimates of the parameters has little effect on the out-of-sample predictive accuracy.

The HB estimates are systematically higher than the modified maximum likelihood estimates across all parameters and data sets. The larger magnitude parameters may indicate smaller error variance in the model, which we would expect from HB estimation, which has the advantage of pooling the data for all respondents. The high proportionality of the parameter estimates across all data sets indicates that both methods will recover similar patterns of heterogeneity.

Table 9 compares the root predictive scores for the holdout tasks for each data set using the MNL, HB, and MML estimation methods. The

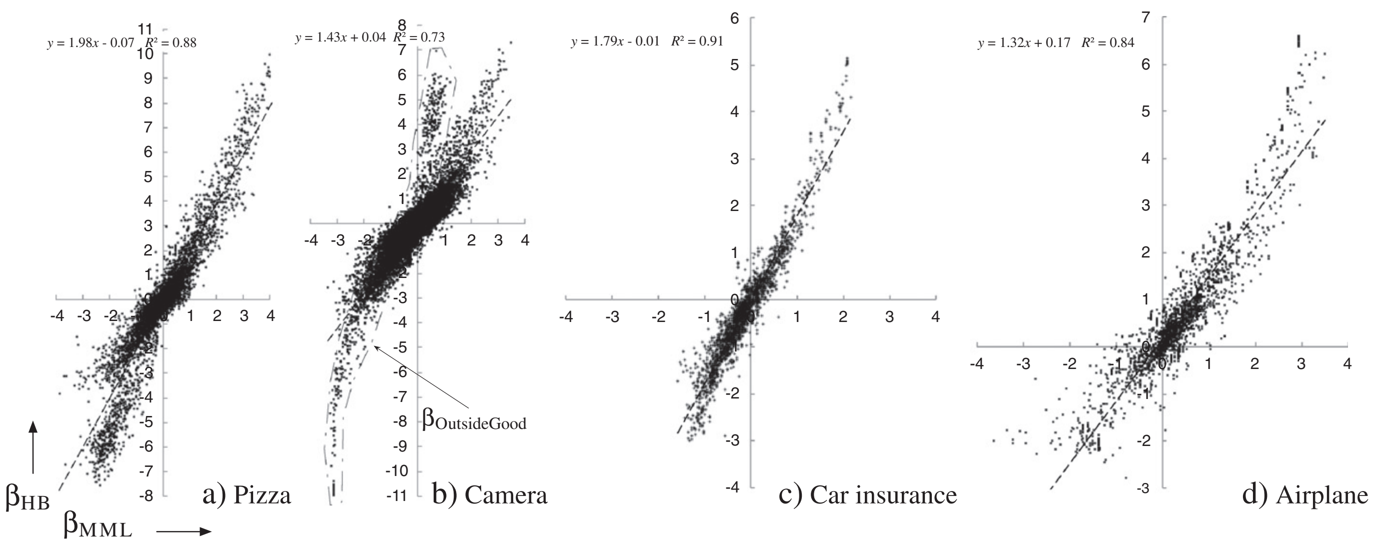

Fig. 5. The modified maximum likelihood estimates, assuming the empirical Bayes prior at the recommended prior weight α_{rec} (x-axis) versus the individual means of the HB posterior estimates (y-axis) for data sets (a) Pizza, (b) Camera, (c) Car insurance, (d) Airline.

Table 10
The number of data separated cases of 1000 simulations for the SAS designs used in Section 3 relative to Bayesian efficient designs assuming a prior similar to the data generating process.

Design	Number of choice tasks				
	4	8	16	32	64
SAS ($\beta = 0$)	1000	995	912	450	30
Bayesian efficient	–	997	846	334	13

results reported in Table 9 are the average root predictive scores $\overline{RPS}$ from 100 validation subsamples of 50% of the full sample and the standard deviations of the *RPS*. The numbers show that all three estimation techniques produce similar root predictive scores for the car insurance and airline data sets. The HB and MML methods produce notably better $\overline{RPS}$ compared to the MNL method for the pizza and camera data sets. The difference in the relative performance of the MNL models between the car insurance and airline data and the pizza and camera data is likely the relative homogeneity in the holdout task responses for the car insurance and airline data sets. For those data sets, the most popular alternatives were chosen by 50–85% of the sample. These were relatively “easy” holdout tasks. In contrast, the most popular alternatives in the pizza and camera data sets were chosen by approximately 30–60% of the sample.

The $\overline{RPS}$ values from the HB method are slightly higher than from the MML method, with the pizza data set showing the largest difference. However, the MML method has smaller standard deviations in three of the four data sets. We would expect HB estimation performance to improve relative to the modified maximum likelihood estimation if the entire predictive distribution from the HB estimation rather than the individual means were used to calculate *RPS*. Also, we expect HB estimation performance to improve as the sample size increases due to the benefits of data pooling.

5. The influence of design

One way to at least partially overcome data separation is through better choice of experiment design. When statistical efficiency is the primary criterion for design selection, a Bayesian efficient design is always preferred to an orthogonal design. Improvements in design efficiency are equivalent to more choice tasks per individual using a less efficient design (Bliemer & Rose, 2011). Therefore, just as data separation is reduced with increasing numbers of choice tasks, data separation is reduced as the efficiency of the design increases. For example, simulations with efficient designs show that the efficient designs result in fewer individuals with separated data (Kessels et al., 2011a,b).

Similarly, Table 10 shows the reduction in the number of simulated individuals with data separation when Bayesian efficient designs with uniform priors are used compared to the orthogonal or nearly orthogonal designs used in Section 3.[10] In general, the new designs lead to a decrease in the number of separated data cases especially for the thirty-two and sixty-four choice task designs. However, there is no improvement for the eight choice task design. The software was not able to improve upon the original design for the four choice task case. This was likely caused by numerical difficulties in evaluating the information matrix for some sets of parameter vectors drawn from the tails of the prior distribution. In all cases, data separation remains a problem with the new designs.

There are three motivations for estimation methods that are robust in the presence of data separation, even given the strides made in efficient design. First, efficient designs take expertise and software (Bliemer & Rose, 2010; Ferrini & Scarpa, 2007; Kessels, Jones, Goos, & Vandebroek, 2009; Kessels et al., 2006; Sándor & Wedel, 2001, 2005; Yu, Goos, & Vandebroek, 2009) that, to date, are not widespread. It is still common practice for many to use readily available orthogonal designs that are more likely to suffer from data separation (Bliemer & Rose, 2011; de Bekker-Grob et al., 2012). Second, the sufficiency of the efficient design approach in eliminating data separation for each and every respondent is rarely known a priori. It is increasingly difficult to eliminate data separation as the number of observations per individual decreases. It is therefore beneficial to have available a method that overcomes difficulties in estimation due to separation, especially when individual parameter estimates are the object. Third, even when data separation is eliminated for a particular individual, the shrinkage induced by the modified maximum likelihood method, or similarly HB estimation, improves individual parameter estimates (Heinze, 2006; King & Ryan, 2002).

6. Conclusion

The traditional way to conceptualize both maximum likelihood estimation and hierarchical Bayes estimation of heterogeneous choice models requires the specification of the functional form of the distribution of customer preferences, although this is not a straightforward or intuitive task for investigators. We demonstrated three alternative approaches that use a computationally simple modified maximum likelihood estimation approach. The first alternative is to construct a notional prior using artificial observations over the choice tasks from a discrete choice experiment. The artificial observations are chosen to represent interpretable prior beliefs about the data, e.g., that all alternatives are equally likely (flat prior). The second approach follows the first, except we assume that a respondent's choices should shrink toward the sample average choices (empirical Bayes prior). The third alternative is to augment the notional sample, representing a flat prior with additional artificial observations over choice tasks that are constructed to be similar to choice tasks in the market of interest (market relevant prior). The artificial observations (i.e., choice probabilities) for these market relevant choice tasks come from an investigator's beliefs or external data.

We presented Monte Carlo simulations and online discrete choice experiment results that demonstrate the method. We explored the impact of varying the weight of the prior on the estimation outcomes both in terms of parameter recovery and prediction accuracy, and we showed that the best value for weight on the notional sample, or prior, depends on the nature of the problem. We detailed a numerical procedure for selecting the prior weight. The results showed that the estimation performance of the proposed modified maximum likelihood method is insensitive over a range of prior weights (identifiable using our numerical procedure) when the performance criterion is either root likelihood for new situations faced by the sample individuals or root predictive score for new situations faced by holdout individuals.

Traditionally, the application of multinomial logistic regression in many economic and marketing problems has been limited to large samples due to data separation in small samples. The data separation checks from the Monte Carlo simulations and the discrete choice experiments indicate that the conventional maximum likelihood estimation of logit models for single individuals and a small number of observations typical of a discrete choice experiment is inadequate. The modified, or penalized, likelihood approach presented in the article overcomes the challenge of data separation and provides an alternative approach for generating individual parameter estimates compared to more complex estimation techniques, such as a hierarchical Bayes method or simulated maximum likelihood method for a random coefficients logit model. In addition to the computational simplicity, the approach enhances the possibility of discrete choice model estimation for very small samples. Small samples are a reality in marketing applications, such as business-to-business products, low volume tourism activities, and products and services involving medical clinicians, to name a few.

[10] Bayesian efficient designs were constructed using Ngene software with 1000 Halton drawings assuming a uniform prior ~U(a, b), where $b - a = 2\sigma$ about a mean μ, where μ and σ are taken from Table 2.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ijresmar.2013.07.005. To access the estimation codes and the data of this paper please refer the following link: www.runmycode.org.

References

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, *71*, 1.

Allenby, G. M., Arora, N., & Ginter, J. L. (1995). Incorporating prior knowledge into the analysis of conjoint studies. *Journal of Marketing Research*, *32*, 152–162.

Allenby, G. M., & Rossi, P. E. (1999). Marketing models of consumer heterogeneity. *Journal of Econometrics*, *89*, 57–78.

Andrews, R. L., Ainslie, A., & Currim, I. S. (2002). An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity. *Journal of Marketing Research*, *39*, 479–487.

Beggs, S., Cardell, N. S., & Hausman, J. (1981). Assessing the potential demand for electric cars. *Journal of Econometrics*, *17*, 1–19.

Bliemer, M. C. J., & Rose, J. M. (2010). Construction of experimental designs for mixed logit models allowing for correlation across choice observations. *Transportation Research Part B: Methodological*, *44*, 720–734.

Bliemer, M. C. J., & Rose, J. M. (2011). Experimental design influences on stated choice outputs: An empirical study in air travel choice. *Transportation Research Part A: Policy and Practice*, *45*, 63–79.

Bull, S. B., Mak, C., & Greenwood, C. M. T. (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis*, *39*, 57–74.

Cardell, N. S. (1993). A modified maximum likelihood estimator for discrete choice models. *Journal of the American Statistical Association: Proceedings of the Statistical Computing Section* (pp. 118–123).

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Chapman, R. G. (1984). An approach to estimating logit models of a single decision maker's choice behavior. *Advances in Consumer Research*, *11*, 656–661.

Chrzan, K., & Orme, B. (2000). An overview and comparison of design strategies for choice-based conjoint analysis. *Technical Paper Research Paper Series Sawtooth Software, Inc.*

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., & Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association*, *86*, 68–78.

de Bekker-Grob, E. W., Ryan, M., & Gerard, K. (2012). Discrete choice experiments in health economics: A review of the literature. *Health Economics*, *21*, 145–172.

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

Ferrini, S., & Scarpa, R. (2007). Designs with a priori information for nonmarket valuation with choice experiments: A Monte Carlo study. *Journal of Environmental Economics and Management*, *53*, 342–363.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*, 27.

Geweke, J. (2005). *Contemporary Bayesian econometrics and statistics*. Hoboken, New Jersey: John Wiley & Sons.

Gilbride, T. J., Lenk, P. J., & Brazell, J.D. (2008). Market share constraints and the loss function in choice-based conjoint analysis. *Marketing Science*, *27*, 995–1011.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*, 359–378.

Haldane, J. B.S. (1955). The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics*, *20*, 309–311.

Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine*, *25*, 4216–4226.

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, *21*, 2409–2419.

James, W., & Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, *Vol. 1*. (pp. 361–379).

Kessels, R., Goos, P., & Vandebroek, M. (2006). A comparison of criteria to design efficient choice experiments. *Journal of Marketing Research*, *43*, 409–419.

Kessels, R., Jones, B., Goos, P., & Vandebroek, M. (2009). An efficient algorithm for constructing Bayesian optimal choice designs. *Journal of Business & Economic Statistics*, *27*, 279–291.

Kessels, R., Jones, B., Goos, P., & Vandebroek, M. (2011a). Rejoinder: The usefulness of Bayesian optimal designs for discrete choice experiments. *Applied Stochastic Models in Business and Industry*, *27*, 197–203.

Kessels, R., Jones, B., Goos, P., & Vandebroek, M. (2011b). The usefulness of Bayesian optimal designs for discrete choice experiments. *Applied Stochastic Models in Business and Industry*, *27*, 173–188.

King, E. N., & Ryan, T. P. (2002). A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. *The American Statistician*, *56*, 163–170.

Kuhfeld, W. F., & Tobias, R. D. (2005). Large factorial designs for product engineering and marketing research applications. *Technometrics*, *47*, 132–141.

Lesaffre, E., & Albert, A. (1989). Partial separation in logistic discrimination. *Journal of the Royal Statistical Society: Series B: Methodological*, *51*, 109–116.

Louviere, J. J., Hensher, D. A., & Swait, J.D. (2000). *Stated choice methods: Analysis and applications*. Cambridge, U.K.: Cambridge University Press.

Mehta, C. R., & Patel, N. R. (1995). Exact logistic regression: Theory and examples. *Statistics in Medicine*, *14*, 2143–2160.

Rossi, P. E., Allenby, G. M., & McCulloch, R. E. (2005). *Bayesian statistics and marketing*. Chichester, West Sussex, UK: John Wiley & Sons.

Sándor, Z., & Wedel, M. (2001). Designing conjoint choice experiments using managers' prior beliefs. *Journal of Marketing Research*, *38*, 430–444.

Sándor, Z., & Wedel, M. (2005). Heterogeneous conjoint choice designs. *Journal of Marketing Research*, *42*, 210–218.

Santner, T. J., & Duffy, D. E. (1986). A note on A. Albert and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, *73*, 755.

Savin, N. E., & Wurtz, A. H. (1999). Power of tests in binary response models. *Econometrica*, *67*, 413–421.

Stein, C. (1956). Inadmissability of the usual estimator for the mean of a multivariate distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, *Vol. 1*. (pp. 197–206).

Train, K. (2003). *Discrete choice methods with simulation*. Cambridge, UK: Cambridge University Press.

Yu, J., Goos, P., & Vandebroek, M. (2009). Efficient conjoint choice designs in the presence of respondent heterogeneity. *Marketing Science*, *28*, 122–135.