Incorporating prior information to overcome complete separation problems in discrete choice model estimation

Bart D. Frischknecht
Centre for the Study of Choice, University of Technology, Sydney, bart.frischknecht@uts.edu.au,

Christine Eckert
Centre for the Study of Choice and Marketing Discipline Group, University of Technology, Sydney, christine.eckert@uts.edu.au,

John Geweke
Centre for the Study of Choice, University of Technology, Sydney, john.geweke@uts.edu.au,

Jordan J. Louviere
Centre for the Study of Choice and Marketing Discipline Group, University of Technology, Sydney, jordan.louviere@uts.edu.au,

We describe a modified maximum likelihood method that overcomes the problem of complete or quasi-complete separation in multinomial logistic regression for discrete choice models with finite samples. The modification consists of augmenting the observed data set with artificial observations that reflect a prior distribution centered on equal choice probabilities. We demonstrate through Monte Carlo simulations of data sets as small the minimum degrees of freedom case that the modified maximum likelihood approach leads to superior parameter recovery and out of sample prediction compared to conventional maximum likelihood. We explore the role the prior weight plays on parameter recovery and out of sample prediction by varying the weight given to the prior versus the weight given to the data in the likelihood function. We demonstrate a numerical procedure to search for an appropriate weight for the prior. One application for the proposed approach is to estimate discrete choice models for single individuals using data from discrete choice experiments. We illustrate this approach with Monte Carlo simulations as well as four data sets collected using online discrete choice experiments.

Key words: choice modeling; maximum likelihood estimation; discrete choice experiment; conjugate prior

1. Introduction

Individuals are heterogeneous. There are many ways to model heterogeneity in quantitative models of human choice, from the simple procedure of estimating a different model for each person, through to much more demanding procedures like setting up a formal distribution of preferences in a population.

A critical drawback in estimating a different model for each person by maximum likelihood is that a single individual's data often exhibits data separation whereby the responses of the individual can be perfectly classified by a linear combination of the covariates. 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. Both complete and quasi-complete separation cases are more likely in small samples, which has been recognized in the biostatistics literature with application to clinical trials (Heinze 2006) and in the econometrics (Beggs et al. 1981) and marketing (Chapman 1984) literatures with previous efforts to estimate choice models using a small sample of data from a single individual.

In the case of complete or quasi-complete data separation (A and Anderson 1984), maximum likelihood estimates for one of the most commonly applied choice models in marketing and economics, the multinomial logistic regression (Train 2003), do not exist. Maximum likelihood estimation in these cases implies that the parameter estimates are unbounded. Although the complete separation problem has been encountered previously in econometrics and marketing (Beggs et al. 1981, Chapman 1984, Savin and Wurtz 1999), it likely has attracted limited attention because of the extensive use of data pooling across respondents leading to large samples and data overlap rather than separation.

Encountering data separation in choice data can be interpreted in one of 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.

We suggest a variant on maximum likelihood estimation that eliminates the separation problem in multinomial logistic regression and greatly improves out-of-sample performance, while imposing

demands on the investigator that are really no greater than doing maximum likelihood person-by-person. Our approach combines various aspects of previous approaches presented in the literature as described in Section 2. Given the assumptions of the multinomial logistic regression model, our method reflects a Bayesian approach, shrinking the values of the parameters towards the implied parameters from the data generating process rather than the parameters tending towards $\pm\infty$.

Our proposed method enables discrete choice model estimation for sample sizes much smaller than most applications reported in literature including the estimation of discrete choice models for single individuals. Applying our modified maximum likelihood approach one respondent at a time to the data of the single respondent, it is possible to construct an empirical distribution of population preferences without prior assumptions of specific population preference distributions and we conjecture with reduced computational burden compared to random coefficients logit estimation either with simulated maximum likelihood or hierarchical Bayes estimation.

Our modified maximum likelihood technique has the advantages of modest computation requirements as it can be implemented in any maximum likelihood estimation software that allows weighted observations, and an intuitive interpretation as the weight given to the prior represents the relative contribution to the model estimation of the prior compared to the observed data. 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 a single individual using the modified maximum likelihood approach and data collected from a stated choice experiment.

The remainder of the article proceeds as follows. Section 2 provides background on the complete separation problem in the context of maximum likelihood estimation and previous efforts to estimate discrete choice models from small samples. Section 3 develops the proposed method for overcoming the complete separation problem in multinomial logistic regression with small samples. Section 4 provides a series of Monte Carlo studies that explore the effect of the proposed method on discrete choice model parameter estimation and prediction. Section 5 illustrates how the approach can be used to estimate a discrete choice model for each individual in a sample and applies this approach to four data sets. Section 6 concludes.

2. Background

A and Anderson (1984) (updated by Santner and Duffy, 1986) identify cases where maximum likelihood estimation will fail to produce binomial logistic regression estimates based on the specific

characteristics of the data. They classify data sets according to the following definitions for complete data separation, quasi-complete data separation, and data overlap: If there exists a vector of parameters $\boldsymbol{\beta}^*$ and corresponding predictor variables $\mathbf{x}_i, i = 1, ..., I$ such that when $\boldsymbol{\beta}^*\mathbf{x}_i > 0, \forall i$ then the observed dependent variable $y_i = 1$ and when $\boldsymbol{\beta}^*\mathbf{x}_i < 0, \forall i$ then $y_i = 0$, the data are called **completely separated**. **Quasi-completely separated** data exhibit the same relations except that the strict inequalities are relaxed such that $\boldsymbol{\beta}^*\mathbf{x}_i = \boldsymbol{\beta}^*\mathbf{x}_j$, when $y_i = 0$ and $y_j = 1$ for at least one i, j pair. All other cases are considered to exhibit **data overlap**.

Maximum likelihood estimates only exist in the case of data overlap. Otherwise the likelihood function will increase as one or more parameters approaches $\pm\infty$. The problem of separation of the data is particularly likely when the number of observations is small, or when a particular alternative(s) is chosen with low probability. The quasi-separated data situation is unlikely when explanatory variables are continuous, but it is plausible in the case of discrete choice experiments where the explanatory variables take a finite set of discrete values. King and Ryan (2002) investigate the case of near separation where data overlap exists but the overlap is in some sense small. They show that for particular underlying parameter values and sample sizes data overlap itself does not guarantee sufficiently small bias as to result in satisfactory estimates.

Various means have been proposed to overcome the data separation challenge especially for biostatistics applications due to small sample sizes and low incidence rates in many clinical trials (Bull et al. 2002, Cardell 1993, Clogg et al. 1991, Firth 1993, Heinze 2006). The proposed approaches including the approach of this article rely on the principle of shrinkage as described by Stein (1956) and James and Stein (1961). The shrinkage is accomplished by estimating parameters based on maximum penalized likelihood. The penalty function adopted corresponds to a Bayesian approach designed to overcome the challenge of finite samples, which is to augment the limited data with prior beliefs about the behavior of the data (Geweke 2005).

The approaches to date can be classified according to the penalty function adopted as either fixed penalty methods or updating penalty methods. The approach of this article belongs to the fixed penalty methods. We first discuss the fixed penalty methods, which add to the data a carefully considered fixed set of artificial observations, thereby assuring data overlap for the extended sample. Haldane (1955), motivated by reducing parameter estimate bias in the binomial logit case, suggests a change to the likelihood formulation for estimation that adds an artificial observation to the data for each binary outcome. Each artificial observation is given half the weight in the log likelihood function as one of the original observations. 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 the binomial case. They consider the relative outcome frequency observed in the data and the number of estimation parameters in determining the number of artificial observations.

The Cardell (1993) approach can be applied to the 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 set observations where J is the total number of unique outcome alternatives (e.g., car, bus, train). The chosen alternative in each artificial choice set is represented by the average of the explanatory variables associated with the alternative 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. This approach appears most appropriate for alternative specific choice models, and even in this case the interpretation of the prior is dependent on the specific values of the explanatory variables.

The fixed penalty methods that add artificial observations to the data are examples of applying a conjugate prior to the data, that is, a prior that has the same distributional form as the likelihood function. In this paper we take a simple approach and imply a flat prior, or probability $\pi_j = 1/J$, for $j = 1, ..., J$ alternatives. The flat prior shrinks parameter estimates towards zeros implying that each alternative is equally likely. It thus pulls the maximum likelihood estimates away from $\pm\infty$ in the case of separated data. An alternative to a flat prior is to adopt an empirical Bayes approach (Carlin and Louis 2000) where the prior implies the aggregate choice shares observed in the sample population. This approach is illustrated briefly in Section 5.1.

Our approach is similar to Clogg et al. (1991) and Cardell (1993) except that Clogg et al. (1991) present their method for the binomial case and for repeated observations of the vectors of explanatory variables only, and the prior employed in Cardell's method is not interpretable for unlabeled-alternative choice models. We present our method for the multinomial case, in choice model format, and for non-repeated vectors of explanatory variables such that our approach is readily applied to data collected from a single individual completing an unlabeled discrete choice experiment.

An alternate yet more complex approach is to derive an updated penalty, that 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. Recently, Kosmidis and Firth (2010) have proposed a model transformation that may facilitate for the multinomial case the iterative computation required to implement the updated penalty method. It is an unresolved issue to consider how one may transform the updated penalty methods for application to discrete choice model estimation.

Apart from fixed and updating penalty methods, exact logistic regression (Mehta and Patel 1995) has been proposed as an alternative to penalized maximum likelihood estimation when data are separated; however, its application is limited in many practical cases given that the method is computationally more intense than penalized maximum likelihood, continuous explanatory variables are not handled well, and confidence intervals can be overly conservative (Heinze 2006). Also, 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.

3. Method

We explain our penalized maximum likelihood method in the context of a data set collected from a discrete choice experiment (Louviere et al. 2000).

3.1. Mathematical description

Assume we have a sample of S completed choice situations, where each choice situation includes J alternatives and where each alternative is described by a $k \times 1$ vector of explanatory variables $\mathbf{x}_{sj}$ where alternatives $j = 1, ..., J$ may or may not be related across choice sets i.e., a labeled or unlabeled design.[1] The dependent variable $y_{sj} = 1$ when alternative j is chosen in choice situation 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

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

Similar to Clogg et al. (1991) we specify a conjugate prior so that the posterior for $\boldsymbol{\beta}$ has the same form as the likelihood function, thus enabling the methodology to be implemented in standard software:

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

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

where g_{sj} are non-negative constants to be specified. When g_{sj} is constrained to be the same value for all s and j, then at the mode of the prior distribution all choices are equally likely for all points $\mathbf{x}_{sj}$ in the sample. We adopt this constraint on the prior for convenience in exposition and therefore drop the subscripts in subsequent notation. The posterior is then

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

The larger is g, the stronger the prior information that choices are equally likely at all sample points $\mathbf{x}_{sj}$. A value of $g = 1$ means that each artificial observation has the same influence on the log likelihood function as each original observation. Given that there are S original observations each with a weight of unity and JS artificial observations each with a weight of g, the ratio R_{prior} of the artificial observations, i.e. the notional prior, to the original data observations is

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

Therefore, values of $g > 1/J$ result in greater weight on the notional prior than on the observed data, and values of $g < 1/J$ result in greater weight on the observed data than on the notional prior.

The prior described in (3) can be implemented in 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 (4).

3.2. Example of method implementation

We show an example in Table 1 when data is to be listed in stacked format, where each row corresponds to a single alternative in a single choice situation. The artificial observations are added to the observed data of S choice situations 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 $\{y_{rj}|r = S+1, ..., S+JS, j = 1, ...J\}$ 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 g_{sj} in the log likelihood function relative to the observed data where the subscript s corresponds to the observed data choice situation for which the artificial choice situation is a replica, and the subscript j corresponds to the alternative j in choice situation s that is chosen in artificial choice situation r, $y_{rj} = 1$. We test the impact of different values of g on the estimation results of the simulation studies and data from four discrete choice experiments.

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

	Choice Situations (S)	Alternative (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	
							g_{sj}	From obs. data choice situation s
Artificial Data	3	1	1	1	1	1	g_{11}	1
	3	2	0	-1	-1	1	g_{11}	1
	3	3	0	-1	-1	-1	g_{11}	1
	4	1	0	1	1	1	g_{12}	1
	4	2	1	-1	-1	1	g_{12}	1
	4	3	0	-1	-1	-1	g_{12}	1
	5	1	0	1	1	1	g_{13}	1
	5	2	0	-1	-1	1	g_{13}	1
	5	3	1	-1	-1	-1	g_{13}	1
	6	1	1	-1	1	-1	g_{21}	2
	6	2	0	-1	-1	1	g_{21}	2
	6	3	0	1	1	-1	g_{21}	2
	7	1	0	-1	1	-1	g_{22}	2
	7	2	1	-1	-1	1	g_{22}	2
	7	3	0	1	1	-1	g_{22}	2
	8	1	0	-1	1	-1	g_{23}	2
	8	2	0	-1	-1	1	g_{23}	2
	8	3	1	1	1	-1	g_{23}	2

4. Method Evaluation Using Monte Carlo Simulations

We first consider the properties of the modified maximum likelihood method using Monte Carlo simulations in a situation typical of the analysis of discrete choice experiments. The Monte Carlo simulations will consider two factors in addition to the weight placed on the prior. The two factors are the distribution from which the simulated data is generated and the number of choice situation observations used in estimation. The performance metrics will be a measure of parameter recovery and a measure of prediction to new choice situations.

The focus of the experiment we study is an object with seven attributes. Attributes $i = 1$ and $i = 2$, which could represent brand and price, each assume one of four levels ($L = 4$). Attributes $i = 3, 4, 5, 6, 7$, representing other features, each assume one of two levels ($L = 2$). For model estimation we use effects coding of the covariates $x_{i\ell},\ \ell = 1, ..., L-1$: if attribute i assumes level $\ell,\ \ell \neq L$ then $x_{i\ell} = 1$ and otherwise $x_{i'\ell'} = 0$; if attribute i assumes level $L,\ x_{i\ell} = -1,\ \ell = 1, ..., L-1$. Utility is linear in attributes:

Table 2 Simulated parameter distribution specifications for the five distributions used in the Monte Carlo Simulation studies

Param.	Normal-small variance†		Normal-large variance		Multivariate normal mixture‡								
	μ_1	σ_1	μ_1	σ_1	Mix.Prob. 1	μ_1	σ_1	Mix.Prob. 2	μ_2	σ_2	Mix.Prob.3	μ_3	σ_3
β_1	0	0.2	0	0.6	0.25	0	0.2	0.45	2	0.1	0.3	0.5	0.7
β_2	-0.4	0.22	-0.4	0.66	0.4	-0.4	0.22	0.15	0.4	0.12	0.45	0.7	0.35
β_3	-0.5	0.2	-0.5	0.6	0.1	-0.5	0.2	0.3	1.5	0.1	0.6	0.2	0.5
β_4	1.9	0.1	1.9	0.3	0.6	1.9	0.1	0.3	2.5	0.1	0.1	0.5	0.18
β_5	-0.6	0.15	-0.6	0.45	0.3	-0.6	0.15	0.5	-0.7	0.05	0.2	0	0.25
β_6	-2.7	0.35	-2.7	1.05	0.2	-2.7	0.35	0.5	-3	0.15	0.3	-0.5	0.2
β_7	0.6	0.12	0.6	0.36	0.35	0.6	0.12	0.25	0	0.02	0.4	1	0.07
β_8	0.5	0.1	0.5	0.3	0.15	0.5	0.1	0.7	2.5	0.2	0.15	1	0.15
β_9	1.4	0.16	1.4	0.48	0.25	1.4	0.16	0.35	-0.4	0.06	0.4	-0.6	0.25
β_{10}	1.9	0.17	1.9	0.51	0.45	1.9	0.17	0.3	0.9	0.07	0.25	0.2	0.45
β_{11}	-1.2	0.13	-1.2	0.39	0.3	-1.2	0.13	0.2	1.8	0.1	0.5	0.4	0.25

†The Fixed parameter distribution has the same parameter means $\boldsymbol{\mu}$ as the Normal distribution but with standard deviations $\boldsymbol{\sigma} = 0$.

‡The Finite mixture parameter distribution has the same parameter means $\boldsymbol{\mu}$ and mixing proportions as the Multivariate normal mixture distribution but with standard deviations $\boldsymbol{\sigma} = 0$.

$$U = \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \beta_4 x_{21} + \beta_5 x_{22} + \beta_6 x_{23} + \beta_7 x_{31} + \beta_8 x_{41} + \beta_9 x_{51} + \beta_{10} x_{61} + \beta_{11} x_{71} + \varepsilon. \quad (6)$$

In each choice situation of the experiment there are four choice alternatives. The respondent's utility is (6), with the random variable ε independently and identically distributed as extreme value type I across all choices and situations. Denoting the 11×1 coefficient vector by $\boldsymbol{\beta}$ and the vector of covariates for choice alternative j of situation s by $\mathbf{x}_{sj}$, the probability P_{sj} of that choice is given by (2).

The Monte Carlo experiment simulates the distribution of the modified maximum likelihood estimate of $\boldsymbol{\beta}$ based on (6) and the choices of a single respondent. The simulated data in the experiment is generated for all combinations of the two Monte Carlo experiment factors: the simulated data distribution and the number of choice situations. For each combination of the two factors in this experiment there are 1,000 simulated data sets.

The first factor for data generation is the population distribution from which the coefficient vector is drawn. There are five alternative distributions: fixed (degenerate), normal small variance, normal large variance, mixture of three fixed, and mixture of three normals. In every case the coefficients are independently but not identically distributed. Table 2 provides the specification of each distribution. We test alternative distributions as a means of testing the modified maximum likelihood method across a broad range of parameter values. Also, in Section 5 we apply the same distributions to study the operationally more relevant problem of predicting population behavior in new choice situations based on models from a subsample of respondents each estimated with the modified maximum likelihood approach.

The second factor for data generation is the number of situations S in which choices are made among four alternatives: $S \in \{4, 8, 16, 32, 64\}$. The design of the five choice experiments composed of different numbers of choice situations, corresponding to the five levels of the second factor for the Monte Carlo simulation experiment, is described as follows.

First, a fractional factorial design is identified as a candidate set of choice alternatives from the full factorial of choice alternatives based on possible combinations of attributes and levels of our problem: $4^2 \times 2^5 = 512$ alternatives. The size of the fractional factorial design chosen depends on the desired number of choice situations. The fractional factorial design consists of 16 alternatives for $S = 4$, 32 alternatives for $S = 8$ and $S = 16$, and 128 alternatives for $S = 32$ and $S = 64$.[2]

Second, the candidate choice alternatives are assigned to choice situations in groups of four. The choice alternatives are assigned to minimize the D-error criterion assuming the means of the parameter values are zero. The D-error is a widely used criterion in the design of choice experiments, and it is defined as the determinant of the parameter variance-covariance matrix (Kessels et al. 2006). We make the assumptions that the means of the parameters are zero $\boldsymbol{\beta} = \mathbf{0}$ and that orthogonal coding is used for the attributes $\mathbf{x}$ when computing the D-error. The result is that the D-error is proportional to the inverse of the information matrix $(\mathbf{x}'\mathbf{x})^{-1}$. The assumption of zero parameter means corresponds to our prior in the modified maximum likelihood formulation that all choice outcomes are equally likely. A modified Fedorov candidate-set-search algorithm implemented in SAS as the %choiceeff macro is used as the search procedure for finding improved candidate designs according to the D-error criterion (Kuhfeld and Tobias 2005).

A single set of T holdout choice situations, identical for all Monte Carlo experimental conditions, is constructed in order to compare the prediction performance in choice situations not used for estimation. The holdout choice situations are constructed as follows. The choice alternatives used for estimation in the choice experiments with choice situations $S \in \{4, 8, 16, 32, 64\}$ are removed from the full factorial of 512 choice alternatives. Due to overlap in the alternatives used in the different sized choice experiments, there remain 357 out of 512 alternatives not assigned to one of the simulation conditions. These 357 alternatives are randomly assigned without replacement in groups of four to $T = 89$ holdout choice situations.

Modified maximum likelihood estimates are constructed for each of the 1,000 data sets in each of the 25 simulation conditions. The simulated responses to the holdout choice situations are also

[2] Fractional factorial designs of other sizes, for example, 64 alternatives for $S = 16$ choice situations, were also considered. Unreported simulations using hierarchical Bayes estimation of a mixed logit model with normally distributed population preferences as well as the modified maximum likelihood estimation indicate that the chosen fractional factorials result in the best parameter recovery on average for the particular simulated data distributions and choice experiment designs tested.

computed and compared with the estimated parameter model predictions for the holdout choice situations. In each case estimation is undertaken for several values of gJ, a ratio of the notional prior data to the observed data. The range of values for gJ is chosen to extend from a near-zero weight prior $gJ = 0.001$ (i.e., the notional prior data contribution is 0.1% of the observed data contribution) to a heavily weighted prior $gJ = 4$ (i.e., the notional prior data contribution is four times the observed data contribution). As the number of choice situations increases, the dependence on the prior is reduced. Therefore, the values of gJ used for a specific simulation condition differ with different numbers of choice situations in order to concentrate the simulation results in the region where the parameter recovery and holdout prediction measures described below attain their best values for this particular problem.

We compute two measures to assess parameter recovery and prediction accuracy across the Monte Carlo experimental factors. We compute the root mean squared error $RMSE$ of each parameter estimate β_k and the estimated parameter $\hat{\beta}_k$ as a measure of parameter recovery (see Andrews et al. (2002) for a similar application of this measure):

$$RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\hat{\beta}_k - \beta_k)^2}, \tag{7}$$

where $K = 11$ is the number of estimated parameters from (6).

The root likelihood is a transformation of the likelihood function that normalizes the likelihood value based on the number of choice situations and can be interpreted as the geometric mean of the model predicted choice probabilities for the observed chosen alternatives. We compute the root likelihood RLH for a series of T holdout choice situations as a measure of prediction accuracy:

$$RLH = L_{HO}^{(1/T)}, \tag{8}$$

where the holdout likelihood L_{HO} is the likelihood function computed as in (1) but for the set of T holdout choice situations rather than the S choice situations used to estimate the model. The prediction measure in this case compares predictions for the sampled population facing T new choice situations.

4.1. 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) discuss identification of data separation in the case of logistic discrimination.

Table 3 Percentage of data sets from Monte Carlo simulation studies with estimated standard deviations that are large in later iterations of traditional maximum likelihood estimation, indicating likely data separation

Number of Choice Sit.	Normal-small variance	Normal-large variance	Fixed	Finite mixture	Multivariate normal mixture
4	99.9%	100%	100%	100%	97%
8	99.7%	100%	100%	100%	100%
16	98.5%	99.1%	98.4%	99.7%	99.6%
32	75.0%	72.6%	74.9%	81.9%	79.2%
64	30.1%	36.7%	29.6%	25.3%	27.4%

They illustrate how large variance estimates at later iterations of traditional maximum likelihood indicate complete or quasi-complete separation rather than collinearity or data overlap. We adopt their finding to perform an ad hoc separation check for 1000 data sets for each of the twenty-five Monte Carlo experimental conditions corresponding to the different parameter distributions and different numbers of choice situations.[3] For each experimental condition Table 3 lists the portion of data sets that returned the standard deviation of at least one parameter as three times larger than the largest (by magnitude) parameter estimate evaluated at the point where the maximum likelihood estimation terminated under standard termination conditions.[4] The percentage of large standard deviations is near 100% for the $S = 4$, $S = 8$, and $S = 16$ choice situation cases. In these cases we expect nearly all data sets to exhibit complete data separation or near complete data separation. The percentage of large standard deviations decreases from 16 to 64 choice situations, yet approximately 30% of data sets still exhibit large standard deviations even for the 64 choice situation case.

4.2. Modified maximum likelihood estimation

First, we illustrate the results from a single data set for a particular set of Monte Carlo simulation experimental factors. A draw from the joint normal-small variance distribution from Table 2 yields parameters $\boldsymbol{\beta}$. Assuming a discrete choice experiment with $S = 16$ choice situations, a notional prior to observed data ratio of $gJ = 0.1$, and a stochastic component $\epsilon_{sj} \sim E.V.1(0, 1)$ added to each alternative j in each choice situation s yields parameter estimates $\hat{\boldsymbol{\beta}}$. $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ for the simulated experiment are plotted in Figure 1. $RMSE$, RLH, $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$ are given in Table 4.

[3] An external check is necessary because our maximum likelihood routine will not identify the estimation problem as unbounded due to satisfaction of one of the numerical convergence criteria such as minimum change in the objective function. This issue was also noted in Beggs et al. (1981) and A and Anderson (1984) for logistic regression and Lesaffre and Albert (1989) for logistic discrimination

[4] The choice of a factor of 3 between the largest standard deviation and the largest parameter estimate is based on experimentation rather than theory.

Figure 1 **The 11 simulated parameters β of (6) from the normal-small variance distribution are plotted against the estimated parameters $\hat{\beta}$ for $S = 16$ choice situations where the estimation has been conducted using the modified maximum likelihood approach with the notional prior to observed data ratio of $gJ = 0.1$.**

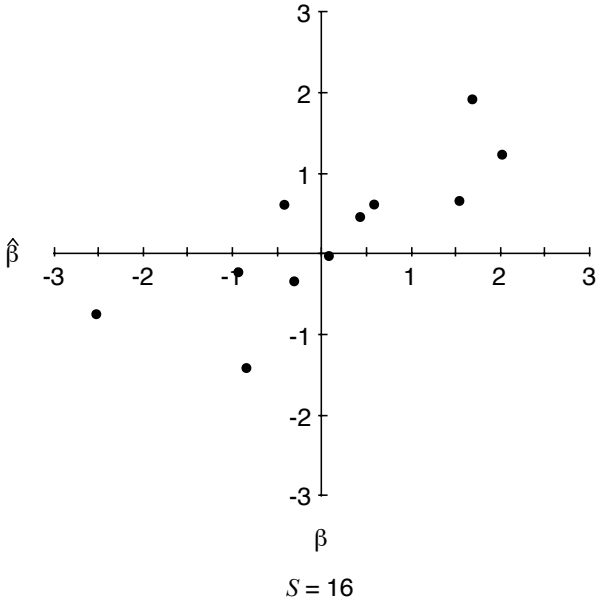
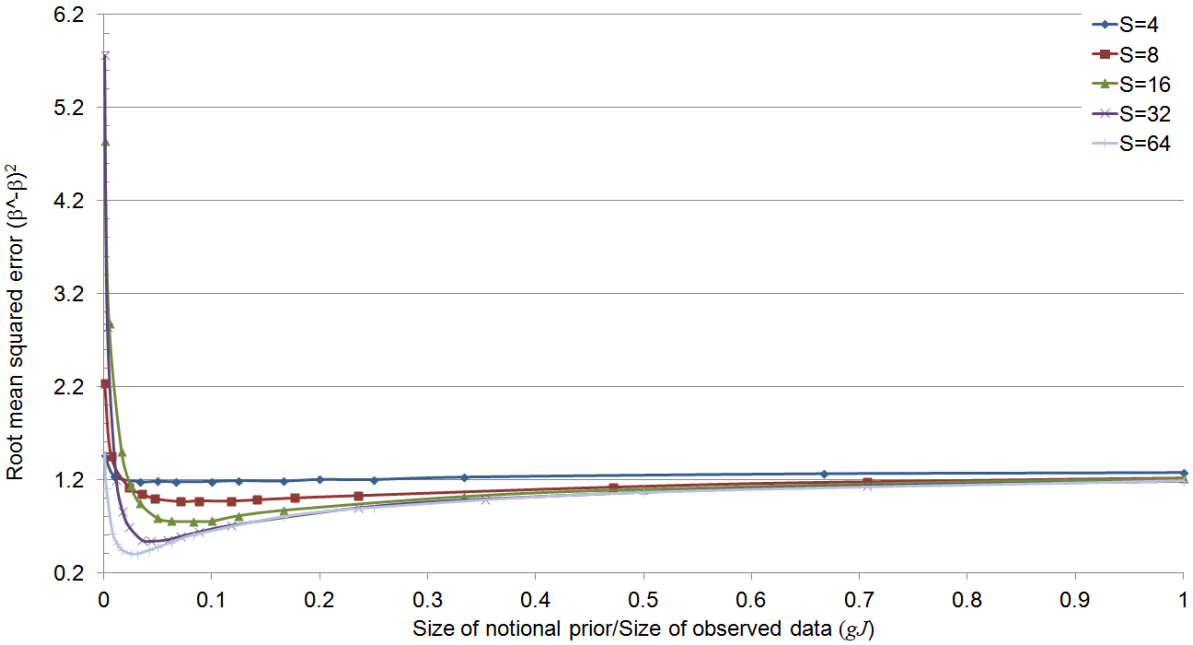


Table 4 **Monte Carlo simulation results for a single data set of simulated preference parameters β from the normal-small variance distribution and the estimated parameters $\hat{\beta}$ for $S = 16$ choice situations where the estimation has been conducted using the modified maximum likelihood approach with the notional prior to observed data ratio of $gJ = 0.1$.**

	$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}}$
β_1	0.08	-0.02
β_2	-0.42	0.61
β_3	-0.31	-0.34
β_4	2.02	1.23
β_5	-0.84	-1.41
β_6	-2.52	-0.75
β_7	0.59	0.62
β_8	0.43	0.46
β_9	1.54	0.66
β_{10}	1.69	1.92
β_{11}	0.93	-0.22
$RMSE$	0.77	
RLH	0.41	

Results such as those in Table 4 are computed for 1000 data sets for each Monte Carlo experimental condition and then averaged for each experimental condition (i.e., an experimental condition consists of a particular parameters distribution, a particular set of choice situations, and a particular prior weight). Figure 2 shows the average root mean squared error between estimated $\hat{\boldsymbol{\beta}}$ and

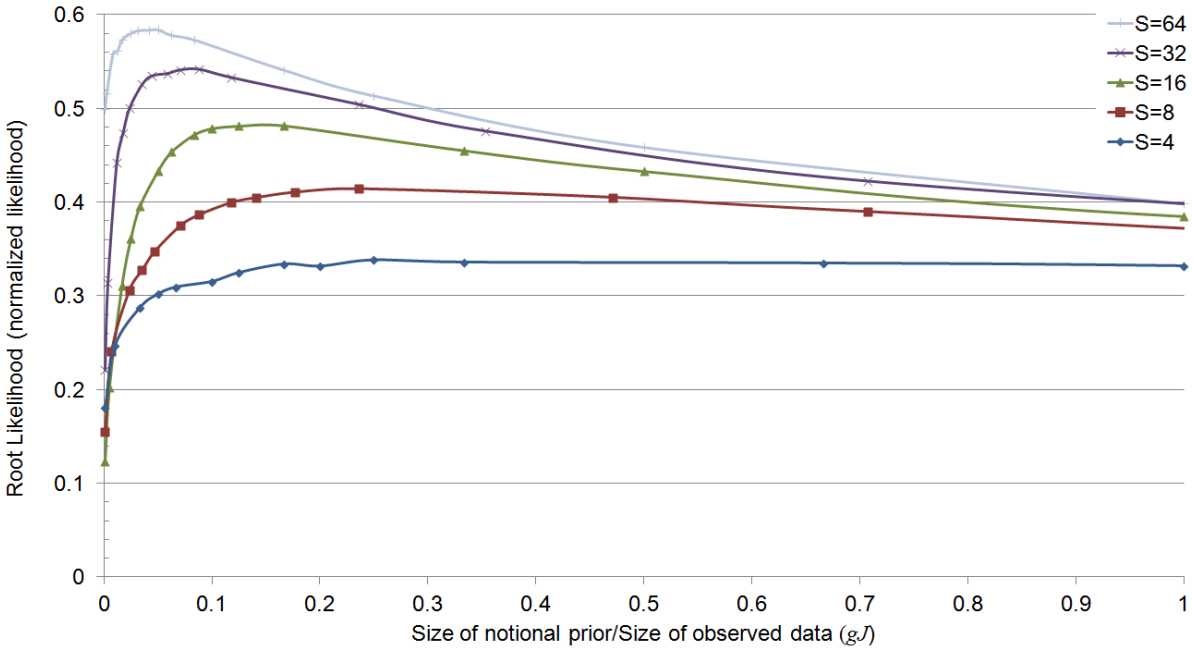
Figure 2 **The average root mean squared errors between the estimated parameters and the simulated parameters averaged over 1000 data sets for each experimental condition as a function of the proportion of artificial data versus observed data used in estimation gJ**



simulated $\boldsymbol{\beta}$ parameters over 1000 data sets for each Monte Carlo experimental condition under the multivariate normal mixture distribution for $0.001 > gJ > 1$. The influence of the prior increases from left to right. Values of $gJ > 1$ indicate that more than half the data used in estimation came from the artificial observations, that is the notional sample supporting the prior that all outcomes are equally likely. The best weight ratio gJ for minimizing root mean squared error increases slightly with decreasing number of choice situations. Figure 2 provides a graphical illustration of the effect of complete or quasi-complete data separation. The curves representing parameter recovery for different numbers of choice situations will be asymptotic to the vertical axis for cases of data separation indicating that root mean squared error between estimated and actual parameters diverges as the weight on the prior goes to zero.

Figure 3 shows the average root likelihood for the holdout choice situations (see (8)) over 1000 data sets for each Monte Carlo experimental condition under the multivariate normal mixture distribution for $0.001 > gJ > 1$. The influence of the prior increases from left to right. The root likelihood is the geometric mean of the predicted choice probabilities for the observed choices of the holdout choice situations, and it is a measure of how the estimated model is performing in predicting choice outcomes in new choice situations. The best weight ratio gJ for maximizing the average root likelihood increases noticeably with decreasing number of choice situations.

Figure 3 **The average root likelihoods for the holdout choice situations averaged over 1000 data sets for each experimental condition as a function of the proportion of artificial data versus observed data used in estimation gJ**



The artificial choice observations are clearly benefiting the estimation as noted by the sharp increase in root mean squared error and sharp decrease in average root likelihood at the left of the graphs as the contribution of the prior to the likelihood approaches zero. The trends with respect to model performance and the trends with respect to weight of the prior are similar for all five parameter distributions. The Appendix provides a table of simulation results for all five parameter distributions. As will be seen in Table 5, the location of the maximum on any one of the curves depends on the parameter values of the distributions in Table 2. As the parameter vector $\boldsymbol{\beta}$ moves closer to zero (where the prior is centered) the best value of the prior weight ratio gJ increases.

5. Specification of prior weight

A typical data set for constructing a discrete choice model for a marketing application would be a set of responses to a number of choice situations from a sample of individuals. Using the estimated model, we would like to predict the choices of a new sample of individuals facing new choice situations. Some choice models assume that population preferences are homogeneous while others model heterogeneity explicitly (Train 2003, Huber and Train 2001). We take the latter approach by estimating a separate choice model for each individual in the sample using the modified maximum likelihood method. We assume that the distribution of sample preferences represents the

Table 5 **The best notional prior to observed data ratio gJ^* and corresponding average root likelihood $R\bar{L}H^*$ for the holdout choice situations averaged over 100 data sets. Results are listed for three different parameter distributions. The distributions are identical to the multivariate normal mixture distribution listed in Table 2 except the means of the parameter distributions have been transformed by a constant product.**

Num of Choice Sit.	$2\boldsymbol{\beta}$		$\boldsymbol{\beta}$		$0.5\boldsymbol{\beta}$	
S	gJ^*	$R\bar{L}H^*$	gJ^*	$R\bar{L}H^*$	gJ^*	$R\bar{L}H^*$
4	0.33	0.36	0.33	0.34	1.0	0.29
8	0.14	0.46	0.24	0.41	0.7	0.33
16	0.10	0.55	0.13	0.48	0.3	0.36
32	0.04	0.65	0.06	0.54	0.24	0.39
64	0.02	0.72	0.04	0.58	0.08	0.42

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.

The modified maximum likelihood method requires the specification of the weighting parameter g to balance the notional prior and the observed data. Whether applied to a single individual or to a sample, the best value for g will vary from problem to problem. As an example, Table 5 compares the values for gJ that maximize average root likelihood performance $R\bar{L}H^*$ for the sample holdout choice situations averaged over 100 data sets for three different problems. As a reminder the root likelihood is the geometric mean of the assigned choice probabilities for the the observed choices in the new choice situations. The data for the three problems were generated from the multivariate normal mixture distribution listed in Table 2 but where the means of the parameter distributions have been transformed by a constant product. Larger parameter values result in choice probabilities close to 0 and 1 while smaller parameters result in choice probabilities closer to equal probability across alternatives. In all cases including the prior helps the estimation in terms of overcoming the data separation problem and achieving higher prediction score, but the value of gJ that yields the largest average root likelihood is smaller as the model parameters are larger in magnitude, i.e., more dissimilar to the prior distribution.

We consider two approaches for specifying the prior weight used in model estimation. First, subjective judgment can provide structure for the problem and an initial weight for the prior. Second, a numerical routine can be implemented to find the prior weight that maximizes model prediction performance for new samples and new choice situations.

5.1. Subjective judgment.

Subjective belief based on experience with the problem should inform the choice of g. For example, a value for g can be chosen based on values of g that have worked well for similar problems.

Additionally, an empirical Bayes approach can be adopted as a method for structuring the prior. Empirical Bayes first uses the data to estimate the prior before using the data again to estimate the model (Carlin and Louis 2000). For example, Clogg et al. (1991) group respondents according to similarity in their covariates. Rather than specify a flat prior with either response outcome expected with $\Pr(y|y=0,1)=0.5$, they take the prior weight as proportional to the frequency of each response category outcome observed across all covariate groups.

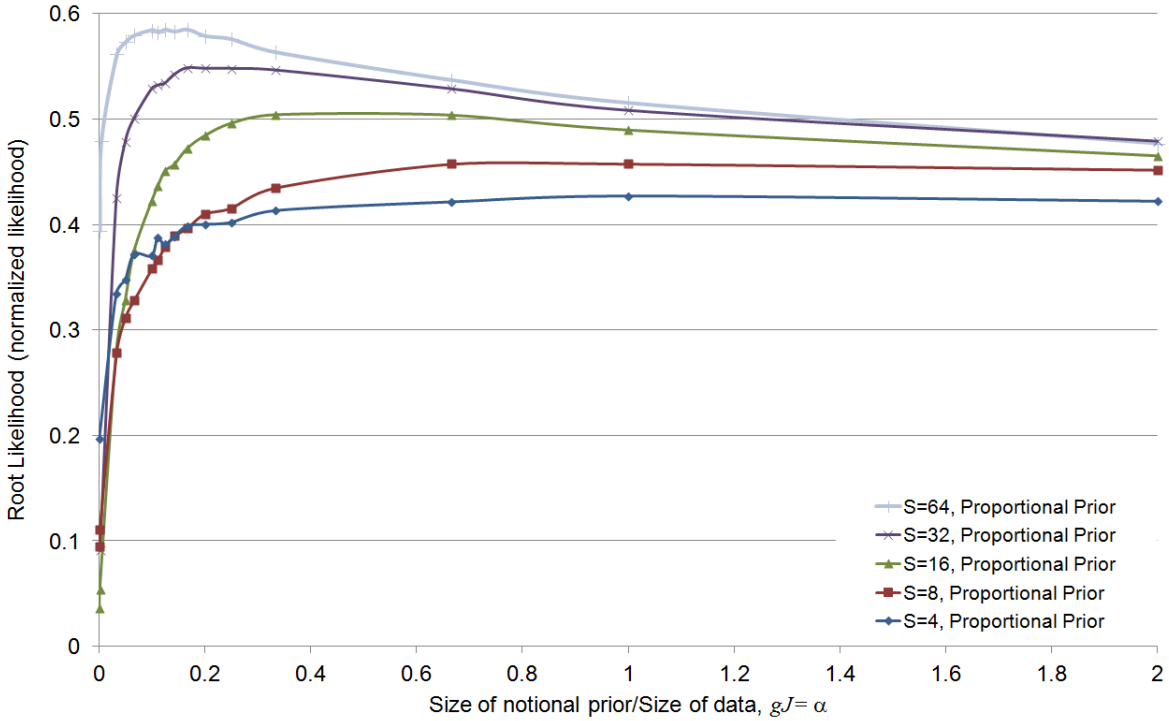
In keeping with this style of empirical Bayes approach, equivalent to shrinking to the observed population mean choice probabilities rather than shrinking to equal probabilities, a similar method can be applied to discrete choice models when a sufficient number of individuals N complete a set of identical choice situations S as in a discrete choice experiment. The sample proportion $\bar{P}_{sj}=1/N\sum_{i=1}^{N}y_{isj}$ of chosen alternative j in choice situation s can be used to modify the prior weights proportionally: $g_{sj}=\alpha\bar{P}_{sj}$, rather than use a constant weight g for all alternatives in all choice situations. The value for α can be set either using subjective beliefs or using a validation sample as described in Section 5.2. For cases when an alternative j in a particular choice situation 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 $\bar{P}_{sj}$.

We apply the sample-proportional choice share prior weight structure ($g_{sj}=\alpha\bar{P}_{sj}$) to the Monte Carlo experiments for 180 data sets (i.e., simulated individuals) drawn from the multivariate normal mixture distribution (Figure 4) and find the average root likelihood for the sampled data sets on the holdout choice situations is larger and the results less sensitive to the choice of prior weight compared to the Monte Carlo experiments that use the equal probability prior (Figure 3). The improvement in average root likelihood on the holdout choice situations diminishes as the number of choice situations per respondent increases so that for $S=64$ choice situations the best average sample root likelihood on the holdout choice situations achieved is the same for the two different prior distributions.

5.2. Numerical prior specification procedure

We seek a measure of model prediction performance that can be calculated for the more general case of predicting the choices of a new sample of individuals facing new choice situations. This is in contrast to the results presented previously where the root likelihood was computed for predictions of the sampled individuals facing new choice situations. The likelihood function as applied in (8) is one example of a scoring function, and many scoring functions can be posited (Gneiting and Raftery 2007). A scoring function is a measure of model prediction performance that relates the probabilistic model predictions to the events that actually occur. We label our measure for scoring

Figure 4 **The average root likelihood for the holdout choice situations for the case of a sample-proportional choice share prior averaged over 100 replications of 180 randomly drawn data sets for each experimental condition as a function of the proportion of artificial data versus observed data used in estimation gJ**



the model on its predictions of the choices T of new individuals M a root predictive score because it is the geometric mean of the predicted choice probabilities π_{itj} based on the model for sample individuals i assigned to the new choice situations t faced by the new individuals m for the choice they are observed to make ($y_{mtj}=1$ when new individual m chooses alternative j in choice situation t and $y_{mtj}=0$ otherwise). The root predictive score is

$$RPS=\left\{\prod_{m=1}^{M}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\prod_{t=1}^{T}\prod_{j=1}^{J}\pi_{itj}^{y_{mtj}}\right)\right]\right\}^{\frac{1}{MT}}, \tag{9}$$

for a population of M new individuals facing T new choice situations each with J alternatives given an estimation sample of N individuals. The following discussion provides a description of (9). Given a collection of estimated models from the sampled individuals and a collection of new individuals, we have no way of knowing the individuals in the estimation sample most similar to a particular new individual from the new sample. We therefore assign equal weight to the models representing the sampled individuals $i=1,...,N$ when predicting the choices of a particular new individual m. The two product terms inside the parentheses represent the product of the assigned probabilities according to the model for sampled individual i for the choice alternatives $j=1,...J$ in choice

situation $t = 1, ..., T$ faced by new individual m. The term $y_{mtj} = 1$ when new individual m chooses alternative j in choice set t and 0 otherwise. The divisor N and the summation $i = 1, ..., N$ inside the brackets represent the averaging of the log score for new individual m when equal weights are given to the models from the sampled individuals $i = 1, ..., N$. The final product term inside the braces indicates that the root predictive score is calculated for the entire sample of new individuals M. The quantity within the braces is raised to the power $1/(MT)$, which results in the root predictive score over new individuals M and new choice situations T. The root predictive score will always take values between 0 and 1 and is a measure of model prediction capability for a new sample facing new choice situations.

The root predictive score RPS (9) is used for defining a procedure to specify a suitable prior weight for modified maximum likelihood estimation. A numerical approach for setting the prior weight g that can be automated with a computer is to conduct the modified maximum likelihood estimation Q times for a particular value of g where for each estimation $q = 1, ..., Q$ a random subsample of half the original sample is withheld from the estimation. The root predictive score for the withheld subsample is calculated according to (9), and the results are averaged over the Q estimations. The procedure is repeated for a finite set of prior weight values g. We recommend selecting values for g such that gJ is in the range $0.001 \leq gJ \leq 2$ with a greater density of trials for the range $0.005 \leq gJ \leq 0.5$. Once a prior weight g has been selected, the models can be reestimated with the entire data set.

As an example Table 6 shows the prior weight g and the average root predictive score measure $\bar{RPS}$ for the withheld subsamples on the $T = 89$ holdout choice situations where the data were generated from the multivariate normal mixture distribution and various numbers of choice situations S. The first set of values gives the value for gJ recommended by the procedure described above and the corresponding average root predictive score. The second set of values gives the best average root predictive score value and the corresponding value for gJ. The final column gives the percentage differences in average root predictive scores based on the best value of gJ and the value of gJ recommend by the prior weight specification procedure.

The prior weight specification procedure achieves average root predictive scores within 1% of the best average root predictive scores observed with the exception of the four choice situation case where the difference is more than 6%. The same trends are exhibited for the data generated from the other distributions in Table 2. The larger value for the best prior weight ratio compared to the value determined by the specification procedure in the case of four choice situations indicates

Table 6 **The prior weight ratio gJ and root predictive score $R\bar{P}S$ for the withheld subsamples on the 89 holdout choice situations where the data were generated from the multivariate normal mixture distribution and various numbers of choice situations S. The first set of values gives the prior weight ratio gJ recommended by the specification procedure and the corresponding $R\bar{P}S$. The second set of values gives the best $R\bar{P}S$ and the corresponding value for gJ. The final column gives the percentage differences in $R\bar{P}S$ based on the best value of gJ and the value of gJ recommend by the prior weight specification procedure.**

Number of Choice Situations	Recommended		Best		Percent difference in RPS between recommended prior and best prior
S	gJ	$R\bar{P}S$	gJ^*	$R\bar{P}S^*$	$(R\bar{P}S - R\bar{P}S^*)/R\bar{P}S^* \times 100$
4	0.07	0.33	0.25	0.35	-6.37
8	0.14	0.43	0.14	0.43	0.00
16	0.10	0.46	0.17	0.47	-0.71
32	0.09	0.48	0.04	0.48	-0.25
64	0.08	0.49	0.05	0.49	-0.24

that the prior weight specification procedure may be inappropriate for cases where the number of choice situations is near the minimum degree of freedom case.[5]

Similar to the results for average root likelihood for the single data sets shown in Figure 3, the average root predictive scores with a near-zero weight prior $gJ = 0.001$ are much lower than the best average root predictive scores observed when more weight is placed on the prior. The large percentage differences between average root predictive scores for the near-zero weight prior and the best prior weight (13-66%) indicate the benefit of the modified maximum likelihood approach and the inadequacy of conventional maximum likelihood for the simulation scenarios. These results combined with the asymptotic relationship observed for both $R\bar{M}SE$ and $R\bar{L}H$ with decreasing weight on the prior as shown in Figures 2-3 offer further support for the use of the modified maximum likelihood approach.

5.3. Discrete Choice Experiment Examples

We test 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 have a sample of 200 respondents and twelve choice situations for estimation and 4 holdout choice situations, and they have the same parameterization as the simulation studies. The pizza data set has a sample of 600 respondents, twenty choice situations for estimation, five holdout choice situations, and it has fifteen parameters corresponding to four four-level attributes, two two-level attributes, and

[5] For discrete choice models, the degrees of freedom in the data are equal to the product of the number of choice situations S and one less than the number of alternatives per choice situation J. The model is said to be saturated when the number of parameters to be estimated is equal to the degrees of freedom in the data.

Table 7 **Description of discrete choice experiments including the percentage of respondents with estimated standard deviations that are large during traditional maximum likelihood estimation, indicating likely data separation**

Data set	Sample Size N	Choice Sit./Resp. S	Alts./Choice Sit. J	HO Choice Sit./ Resp. S_{HO}	Num. of Params. K	% Large Std. Dev.
Pizza	600	20	5	5	15	99.8
Camera	600	24	5	5	25	99.7
Car Insurance	200	12	4	4	11	96.5
Airline	200	12	4	4	11	94.0

a no-choice alternative constant. The digital camera data set has a sample of 600 respondents, twenty-four choice situations for estimation, five holdout choice situations, and it has 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 no-choice alternative constant. Table 7 summarizes the discrete choice experiments including the percentage of respondents that appear to have parameter estimates with large standard deviations when the termination criterion is reached in conventional maximum likelihood estimation. As explained in Section 4.1, we take the large standard deviations of the parameter estimates in later iterations of conventional maximum likelihood to indicate likely data separation or near data separation.

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 5.2 to identify a suitable prior weight, first using the equal probability prior and then using the sample-proportional prior described in Section 5.1. The online discrete choice experiment results exhibit the same trends with respect to changes in gJ as the simulation results. Similar to Table 6, we report in Table 8 the notional prior to observed data ratio gJ recommended by the weight specification procedure and the resulting average root predictive score $\bar{RPS}$ as well as the best weight ratio gJ^* and average root predictive score $\bar{RPS}^*$ for each data set. The percentage differences between the best average root predictive scores and the average root predictive scores achieved by the prior weight specification procedure are between 0-2%. The best value for the notional prior weight ratio 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 weight ratios are between 0.12-0.50.

The prior weight specification 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 the benefit of the modified maximum likelihood approach for individual model estimation from the four online discrete choice experiments.

Table 8 **The prior weight ratio gJ and average root predictive score $R\bar{P}S$ for the withheld subsamples on the holdout choice situations where the data came from online discrete choice experiments. The first set of values gives results for the equal probability prior, and the second set of results gives the results for the sample-proportional probability prior. The results are listed from left to right as the prior weight ratio gJ recommended by the specification procedure and the corresponding root predictive score; the best root likelihood value $R\bar{P}S^*$ and the corresponding value for gJ^*; the percentage differences in root predictive scores based on the best value of gJ^* and the value of gJ recommend by the prior weight specification procedure.**

Data Set	Equal probability prior					Sample-proportional probability prior				
	Recommended		Best		Percent difference in $R\bar{P}S$ between recommended prior and best prior	Recommended		Best		Percent difference in $R\bar{P}S$ between recommended prior and best prior
	gJ	$R\bar{P}S$	gJ^*	$R\bar{P}S^*$	$\frac{R\bar{P}S - R\bar{P}S^*}{R\bar{P}S^*} \times 100$	gJ	$R\bar{P}S$	gJ^*	$R\bar{P}S^*$	$\frac{R\bar{P}S - R\bar{P}S^*}{R\bar{P}S^*} \times 100$
Pizza	0.17	0.278	0.07	0.284	-2.04%	0.17	0.282	0.10	0.286	-1.52%
Camera	0.30	0.241	0.15	0.243	-0.77%	0.33	0.242	0.25	0.244	-0.63%
Car Insurance	0.19	0.380	0.06	0.384	-0.99%	0.50	0.394	0.67	0.400	-1.69%
Airline	0.12	0.380	0.02	0.386	-1.58%	0.25	0.395	0.03	0.395	-0.03%

Computation time for the method will be proportional to the number of modified maximum likelihood estimations conducted. This means that computational time is proportional to sample size, the number of sample/validation subsample pairs, and the number of values of gJ tested. The number of modified maximum likelihood estimations 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., 10), and the number of values of gJ tested (i.e., 15). The other significant computation is the calculation of the root predictive score, which is calculated for each sample/validation subsample and each value of gJ tested. We used 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 take 25 seconds. A single calculation of the root predictive score takes 0.62 seconds. This means the total computation time of the modified maximum likelihood estimations of the pizza data took about 64 minutes. Given the relative insensitivity of root predictive score with respect to gJ near the best value of $R\bar{P}S$ it is plausible to consider only a single estimation/validation subsample pair and fewer test values for gJ. This would lead to a reduction in computation time of two orders of magnitude below what was achieved for the results in Table 8. Further computation time reduction can be achieved by estimating the individual models in parallel rather than in series.

6. Conclusion

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. Analysis progressed historically from aggregate homogeneous models to finite mixture models and to random coefficients models. This article presents a penalized likelihood approach that provides an alternative approach for generating individual parameter estimates compared to more complex estimation techniques such as simulated maximum likelihood for a random coefficients logit model. The approach also opens the possibility of discrete choice model estimation for small samples. Small samples are a reality in marketing applications such as business-to-business products, low volume tourism activities, and products involving medical clinicians, to name a few.

Many data generating processes can result in complete data separation when observations are limited and the variable space is complex including choice behavior that conforms to a multinomial logistic regression interpretation. Our approach is similar to approaches proposed in the statistics literature for binary logistic regression and alternative forms of multinomial logistic regression. We illustrate our approach specifically for multinomial logistic regression with data in a format typical to discrete choice experiments.

We present Monte Carlo simulation and online discrete choice experiment data estimated results for the multinomial case. We explore the impact of varying the weight of the prior on the estimation outcomes both in terms of parameter recovery and prediction accuracy, and we show that the best value for weight on the penalty function, or prior, depends on the nature of the problem. We detail a numerical procedure for selecting the prior weight for a specific case.

The frequent occurrence of data separation in small samples and the Monte Carlo simulations and the discrete choice experiment results with near zero weight on the prior indicate that estimating conventional maximum likelihood estimation of logit models for single individuals and a small number of observations typical of a discrete choice experiment is completely inadequate. The results show that the proposed modified maximum likelihood shows little variation in performance over a range of prior weights 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 new individuals.

Future work should explore the relationship between the approach taken in this article and the updating penalty methods proposed by Firth (1993) and Kosmidis and Firth (2010). Also, the individual estimation approach described here using modified maximum likelihood should be compared to pooled data estimation techniques such as hierarchical Bayes.

Appendix. Monte Carlo Simulation Results

Monte Carlo simulation results for the average sample root means squared error $R\bar{M}SE$ between the estimated parameters and the simulated parameters and the average sample root likelihood for the hold-out choice situations $R\bar{L}H$ averaged over 1000 data sets from the parameter joint distribution for each experimental condition are listed in Table 9.

References

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

Andrews, R. L., A. Ainslie, I. S. Currim. 2002. An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity. *Journal of Marketing Research* **39**(4) 479–487.

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

Bull, S. B., C. Mak, C. M. T. Greenwood. 2002. A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics & Data Analysis* **39**(1) 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*. 118–123.

Carlin, B. P., T. A. Louis. 2000. *Bayes and Empirical Bayes Methods for Data Analysis*. 2nd ed. Chapman and Hall/CRC, Boca Raton, FL.

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.

Clogg, C. C., D. B. Rubin, N. Schenker, B. Schultz, L. Weidman. 1991. Multiple imputation of industry and occupation codes in census public-use samples using bayesian logistic regression. *Journal of the American Statistical Association* **86**(413) 68–78.

Firth, D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika* **80**(1) 27.

Geweke, J. 2005. *Contemporary Bayesian Econometrics and Statistics*. John Wiley & Sons, Hoboken, New Jersey.

Gneiting, T., A. E. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**(477) 359–378.

Haldane, J. B. S. 1955. The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics* **20**(4) 309–311.

Heinze, G. 2006. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in medicine* **25**(24) 4216–4226.

Heinze, G., M. Schemper. 2002. A solution to the problem of separation in logistic regression. *Statistics in medicine* **21**(16) 2409–2419.

Huber, J., K. Train. 2001. On the similarity of classical and bayesian estimates of individual mean partworths. *Marketing Letters* **12**(3) 259–269.

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

Kessels, R., P. Goos, M. Vandebroek. 2006. A comparison of criteria to design efficient choice experiments. *Journal of Marketing Research* **43**(3) 409–419.

King, E. N., T. P. Ryan. 2002. A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. *The American Statistician* **56**(3) pp. 163–170.

Kosmidis, I., D. Firth. 2010. Multinomial logit bias reduction via poisson log-linear model. Tech. Rep. CRiSM 10-18, University of Warwick.

Kuhfeld, W. F., R. D. Tobias. 2005. Large factorial designs for product engineering and marketing research applications. *Technometrics* **47**(2) 132–141.

Lesaffre, E., A. Albert. 1989. Partial separation in logistic discrimination. *Journal of the Royal Statistical Society.Series B (Methodological)* **51**(1) pp. 109–116.

Louviere, J. J., D. A. Hensher, J. D. Swait. 2000. *Stated Choice Methods: Analysis and Applications*. Cambridge University Press, Cambridge, U.K.

Mehta, C. R., N. R. Patel. 1995. Exact logistic regression: theory and examples. *Statistics in medicine* **14**(19) 2143–2160.

Santner, T. J., D. E. Duffy. 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**(3) 755.

Savin, N. E., A. H. Wurtz. 1999. Power of tests in binary response models. *Econometrica* **67**(2) pp. 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. 197–206.

Train, K. 2003. *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge, UK.

Table 9 **The Monte Carlo experiment results averaged over 1000 data sets for each experimental condition tested. The first two columns list the number of choice situations and the prior weight ratio. The first column for each distribution lists the average sample root mean squared error between the estimated parameters and the simulated parameters. The second column lists average sample root likelihood for the holdout choice situations**

Num of Choice Sit.	Weight	Normal-small variance		Normal-large variance		Fixed		Finite mixture		Multivariate normal mixture	
S	gJ	$RM\bar{S}E$	$R\bar{L}H$	$RM\bar{S}E$	$R\bar{L}H$	$RM\bar{S}E$	$R\bar{L}H$	$RM\bar{S}E$	$R\bar{L}H$	$RM\bar{S}E$	$R\bar{L}H$
4	4.000	1.30	0.28	1.37	0.28	1.29	0.28	1.35	0.29	1.36	0.29
4	2.000	1.28	0.29	1.35	0.30	1.27	0.29	1.30	0.32	1.33	0.32
4	1.000	1.26	0.29	1.32	0.30	1.25	0.29	1.27	0.33	1.27	0.33
4	0.667	1.25	0.28	1.32	0.30	1.25	0.28	1.24	0.34	1.26	0.33
4	0.333	1.26	0.27	1.30	0.29	1.24	0.26	1.20	0.34	1.22	0.34
4	0.250	1.25	0.26	1.30	0.28	1.25	0.25	1.18	0.34	1.20	0.34
4	0.200	1.25	0.25	1.29	0.27	1.25	0.25	1.18	0.33	1.20	0.33
4	0.167	1.27	0.24	1.30	0.27	1.26	0.24	1.17	0.33	1.18	0.33
4	0.125	1.28	0.23	1.30	0.26	1.27	0.23	1.17	0.32	1.19	0.32
4	0.100	1.29	0.22	1.31	0.25	1.28	0.22	1.16	0.32	1.18	0.32
4	0.067	1.31	0.20	1.31	0.23	1.30	0.20	1.16	0.31	1.17	0.31
4	0.050	1.32	0.19	1.34	0.22	1.31	0.19	1.16	0.30	1.18	0.30
4	0.033	1.35	0.17	1.36	0.20	1.36	0.17	1.17	0.28	1.18	0.29
4	0.010	1.49	0.13	1.48	0.15	1.50	0.12	1.22	0.25	1.24	0.25
4	0.001	1.77	0.07	1.77	0.09	1.79	0.07	1.45	0.18	1.46	0.18
8	2.828	1.24	0.31	1.32	0.32	1.23	0.31	1.30	0.32	1.33	0.32
8	1.414	1.17	0.35	1.24	0.35	1.16	0.35	1.25	0.35	1.25	0.35
8	0.707	1.09	0.38	1.17	0.39	1.08	0.38	1.16	0.39	1.17	0.39
8	0.471	1.02	0.40	1.11	0.40	1.01	0.40	1.11	0.40	1.12	0.40
8	0.236	0.91	0.41	0.99	0.42	0.91	0.41	1.01	0.42	1.03	0.41
8	0.177	0.87	0.41	0.97	0.42	0.86	0.41	0.98	0.41	1.00	0.41
8	0.141	0.85	0.41	0.91	0.41	0.84	0.41	0.96	0.41	0.98	0.40
8	0.118	0.83	0.40	0.91	0.40	0.82	0.40	0.95	0.40	0.97	0.40
8	0.088	0.83	0.39	0.92	0.39	0.81	0.39	0.95	0.39	0.97	0.39
8	0.071	0.85	0.38	0.91	0.38	0.84	0.37	0.96	0.37	0.96	0.37
8	0.047	0.88	0.36	0.95	0.36	0.88	0.35	1.00	0.35	0.99	0.35
8	0.035	0.91	0.34	1.00	0.35	0.91	0.34	1.07	0.32	1.05	0.33
8	0.024	1.01	0.32	1.09	0.32	1.00	0.32	1.17	0.30	1.12	0.31
8	0.007	1.37	0.26	1.46	0.26	1.37	0.26	1.52	0.23	1.45	0.24
8	0.001	2.11	0.18	2.29	0.18	2.11	0.18	2.24	0.15	2.23	0.16
16	2.000	1.20	0.34	1.27	0.34	1.18	0.34	1.26	0.34	1.29	0.34
16	1.000	1.11	0.39	1.19	0.39	1.10	0.39	1.18	0.38	1.21	0.38
16	0.500	0.99	0.44	1.07	0.44	0.98	0.44	1.08	0.43	1.09	0.43
16	0.333	0.91	0.46	1.01	0.47	0.90	0.46	1.00	0.45	1.02	0.45
16	0.167	0.75	0.48	0.85	0.49	0.75	0.48	0.85	0.48	0.87	0.48
16	0.125	0.70	0.48	0.78	0.49	0.69	0.48	0.80	0.48	0.81	0.48
16	0.100	0.65	0.48	0.73	0.49	0.65	0.48	0.77	0.47	0.75	0.48
16	0.083	0.65	0.47	0.72	0.49	0.63	0.47	0.75	0.46	0.75	0.47
16	0.063	0.67	0.45	0.71	0.47	0.64	0.45	0.74	0.45	0.75	0.45
16	0.050	0.70	0.44	0.76	0.45	0.69	0.43	0.79	0.43	0.78	0.43
16	0.033	0.91	0.40	0.91	0.42	0.90	0.39	0.95	0.39	0.94	0.40
16	0.025	1.14	0.36	1.12	0.38	1.17	0.35	1.14	0.36	1.13	0.36
16	0.017	1.55	0.30	1.48	0.33	1.54	0.30	1.54	0.31	1.50	0.31
16	0.005	3.03	0.19	2.97	0.22	3.18	0.18	2.85	0.20	2.88	0.20
16	0.001	5.12	0.12	4.90	0.14	5.48	0.11	4.86	0.13	4.84	0.12
32	1.414	1.13	0.37	1.22	0.37	1.12	0.37	1.21	0.37	1.24	0.37
32	0.707	1.02	0.42	1.09	0.43	1.00	0.42	1.11	0.42	1.13	0.42
32	0.354	0.87	0.48	0.95	0.49	0.86	0.47	0.97	0.47	0.99	0.48
32	0.236	0.77	0.50	0.86	0.51	0.77	0.50	0.88	0.50	0.89	0.50
32	0.118	0.60	0.53	0.68	0.54	0.59	0.53	0.70	0.53	0.70	0.53
32	0.088	0.53	0.53	0.61	0.55	0.52	0.53	0.63	0.54	0.63	0.54
32	0.071	0.49	0.53	0.56	0.55	0.48	0.53	0.58	0.54	0.58	0.54
32	0.059	0.48	0.53	0.54	0.55	0.47	0.53	0.54	0.54	0.55	0.54
32	0.044	0.50	0.52	0.53	0.54	0.49	0.52	0.53	0.53	0.53	0.53
32	0.035	0.55	0.51	0.56	0.53	0.55	0.52	0.54	0.53	0.55	0.53
32	0.024	0.74	0.49	0.71	0.50	0.74	0.49	0.67	0.50	0.69	0.50
32	0.018	0.95	0.46	0.93	0.48	0.95	0.46	0.88	0.47	0.85	0.47
32	0.012	1.28	0.43	1.28	0.44	1.31	0.42	1.19	0.44	1.19	0.44
32	0.004	3.00	0.30	3.00	0.32	2.90	0.30	2.82	0.31	2.84	0.31
32	0.001	5.58	0.24	5.83	0.23	5.53	0.23	5.56	0.23	5.76	0.22
64	1.000	1.07	0.41	1.17	0.41	1.06	0.41	1.16	0.40	1.18	0.40
64	0.500	0.94	0.47	1.04	0.47	0.93	0.47	1.04	0.46	1.06	0.46
64	0.250	0.78	0.52	0.88	0.52	0.76	0.52	0.88	0.51	0.90	0.51
64	0.167	0.68	0.54	0.77	0.55	0.66	0.54	0.77	0.54	0.80	0.54
64	0.083	0.50	0.57	0.59	0.58	0.49	0.56	0.58	0.57	0.60	0.57
64	0.063	0.44	0.57	0.51	0.58	0.43	0.57	0.51	0.58	0.52	0.58
64	0.050	0.40	0.57	0.46	0.59	0.39	0.57	0.46	0.58	0.47	0.58
64	0.042	0.38	0.57	0.43	0.59	0.38	0.57	0.42	0.58	0.43	0.58
64	0.031	0.38	0.57	0.41	0.59	0.38	0.56	0.39	0.58	0.40	0.58
64	0.025	0.39	0.57	0.41	0.58	0.40	0.56	0.39	0.58	0.40	0.58
64	0.017	0.48	0.55	0.47	0.57	0.48	0.55	0.43	0.57	0.44	0.57
64	0.013	0.55	0.54	0.54	0.56	0.55	0.54	0.51	0.56	0.51	0.56
64	0.008	0.68	0.53	0.68	0.55	0.70	0.53	0.62	0.55	0.62	0.56
64	0.003	1.09	0.49	1.16	0.51	1.06	0.49	1.08	0.52	1.11	0.52
64	0.001	1.38	0.47	1.67	0.48	1.55	0.45	1.44	0.50	1.47	0.50