Statistica Sinica **19** (2009), 731-748

EXISTENCE OF THE MLE AND PROPRIETY OF POSTERIORS FOR A GENERAL MULTINOMIAL CHOICE MODEL

Paul L. Speckman[1], Jaeyong Lee[2] and Dongchu Sun[1]

[1] *University of Missouri-Columbia and* [2] *Seoul National University*

Abstract: This paper examines necessary and sufficient conditions for the existence of Maximum Likelihood Estimates (MLE) and the propriety of the posterior under a bounded improper prior density for a wide class of discrete (or multinomial) choice models. The choice models are based on the principle of utility maximization. Our results cover a wide class of latent variable distributions defining the utility, including in particular multinomial logistic and probit classification and choice models as special cases. Albert and Anderson (1984) gave separation and overlap conditions for the existence of the MLE in logistic classification models. We generalize their conditions to multinomial choice models, giving necessary and sufficient conditions for the existence of a finite MLE and the propriety of the posterior for a wide class of bounded improper priors. Consistency and asymptotic normality for both the MLE and the posterior are also proved under mild conditions.

Key words and phrases: Asymptotic normality, logistic, maximimum likelihood estimation, multinomial choice model, posterior, probit.

1. Introduction

Bayesian inference in discrete choice models such as logistic and probit regression has received a great deal of recent attention. While many Bayesians favor using proper subjective priors, in practice there may be limited information or time constraints that make the use of noninformative or default priors desirable. Other statisticians prefer noninformative priors on philosophical grounds. Raghavan and Cox (1998) considered noninformative priors with a bounded prior density for the binary logistic model, and Albert and Chib (1993) proposed a constant prior for the binary probit model. A natural question is whether the posterior is proper for a general multinomial choice model if a bounded improper prior, such as a constant prior, is used. Without proper precaution, simple noninformative priors such as a constant prior can be misused, sometimes unknowingly (for example, see Hobert and Casella (1996)).

We have been interested in both classical and Bayesian inference for a general class of discrete choice models (including logistic and probit regression). This

interest has led us to investigate some fundamental issues regarding existence of estimators. Although the behavior of maximum likelihood in the usual logistic models is well understood, surprisingly there seems to be a gap in the literature for more general models. A notable exception is Chen, Ibrahim and Shao (2004), who considered the propriety of the posterior in generalized linear models when covariates are missing.

The purpose of this paper is to study the existence of maximum likelihood estimates and the propriety of the posterior when an improper prior is used. Our main results show that, in a broad class of problems, the existence of the two estimators coincide. In addition, we demonstrate asymptotic normality of the MLE and the posterior under mild conditions.

The classes of models considered here have the following properties. Suppose $y_1, \ldots, y_n$ are independent random variables, where

$$y_i \sim \text{ multinomial}(1, \boldsymbol{p}(\boldsymbol{X}_i, \boldsymbol{\beta})) \text{ for } i = 1, \ldots, n, \tag{1.1}$$

$\boldsymbol{p}(\boldsymbol{X}_i, \boldsymbol{\beta}) = (p_1(\boldsymbol{X}_i, \boldsymbol{\beta}), \ldots, p_k(\boldsymbol{X}_i, \boldsymbol{\beta}))^t$, $P(y_i = j \mid \boldsymbol{\beta}) = p_j(\boldsymbol{X}_i, \boldsymbol{\beta})$, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{ik})$ is an $m \times k$ design matrix for the ith observation, $\boldsymbol{\beta}$ is a parameter vector of length m, and $\boldsymbol{p}(\boldsymbol{X}_i, \boldsymbol{\beta})$ satisfies

$$p_j(\boldsymbol{X}_i, \boldsymbol{\beta}) \geq 0 \text{ for } j = 1, \ldots, m; \quad \sum_{j=1}^{k} p_j(\boldsymbol{X}_i, \boldsymbol{\beta}) = 1.$$

There are two common forms of parameterization for these models, which we term *choice models* and *classification models*. References to choice models go back to Thurstone (1927), Luce (1959), etc., in psychology. Multinomial choice models have a long history in economics and transportation, see e.g., Anas (1983), Ben-Akiva and Lerman (1985) and Anderson, de Palma and Thisse (1992). As a simple example, consider the following model for choosing a location to shop. Person i has k available shopping location choices. Each location has properties that make it more or less attractive as a shopping destination: the distance from person i's home, a measure of available shopping opportunities such as the number of retail employees at the location, and possible interaction with the socio-economic status of person i. These variables are captured in a vector of covariates $\boldsymbol{x}_{ij}$ of length m and depend in general on both characteristics of person i and destination j. In the multinomial logistic model, the probability that person i makes choice j is

$$p_j(\boldsymbol{X}_i, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_{ij}^t \boldsymbol{\beta})}{\exp(\boldsymbol{x}_{i1}^t \boldsymbol{\beta}) + \cdots + \exp(\boldsymbol{x}_{ik}^t \boldsymbol{\beta})}, \tag{1.2}$$

where $\boldsymbol{\beta}$ is an unknown vector with length m. Note that the specification is not unique. For example, $\boldsymbol{x}_{ij}$ can be replaced by $\boldsymbol{x}_{ij} - \boldsymbol{x}_{ik}$, for $j = 1, \ldots, k-1$, with $\boldsymbol{x}_{ik} = 0$.

The second model is one we term the classification model. To illustrate, suppose patient i is to be classified into one of k disease states on the basis of a vector of measurements $\boldsymbol{x}_i$ of length r. A common model is the logistic model,

$$P(y_i = j \mid \boldsymbol{\beta}) = \begin{cases} \frac{\exp(\boldsymbol{x}_i^t \boldsymbol{\beta}_j)}{1+\sum_{l=1}^{k-1} \exp(\boldsymbol{x}_i^t \boldsymbol{\beta}_l)}, & j = 1, \ldots, k-1, \\ \frac{1}{1+\sum_{l=1}^{k-1} \exp(\boldsymbol{x}_i^t \boldsymbol{\beta}_l)}, & j = k, \end{cases} \tag{1.3}$$

for parameter vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{k-1}$. The binary case is most common. If $k = 2$, we assume that $y_i = 0$ or $y_i = 1$, and $P(y_i = 1 \mid \boldsymbol{\beta}) = \exp(\boldsymbol{x}_i^t \boldsymbol{\beta})/\{1+\exp(\boldsymbol{x}_i^t \boldsymbol{\beta})\} = 1 - P(y_i = 0 \mid \boldsymbol{\beta})$. In this case, the choice and classification models are equivalent. In general, the classification model is a special case of the choice model. Let $m = r(k-1)$, and define $\boldsymbol{x}_{i1}^t = (\boldsymbol{x}_i^t, \mathbf{0}^t, \ldots, \mathbf{0}^t), \ldots, \boldsymbol{x}_{i,k-1}^t = (\mathbf{0}^t, \ldots, \mathbf{0}^t, \boldsymbol{x}_i^t)$, $\boldsymbol{x}_{ik}^t = (\mathbf{0}^t, \ldots, \mathbf{0}^t)$. With $\boldsymbol{\beta}^t = (\boldsymbol{\beta}_1^t, \ldots, \boldsymbol{\beta}_{k-1}^t)$, the classification model (1.3) has exactly the form (1.2).

These choice models can be motivated and generalized using the principle of utility maximization. Suppose there are continuous latent random variables ξ_{ij} such that $\boldsymbol{x}_{ij}^t \boldsymbol{\beta} + \xi_{ij}$ is the utility of choice j to person i. Assuming that individuals act to maximize utility, person i makes choice j if

$$\boldsymbol{x}_{ij}^t \boldsymbol{\beta} + \xi_{ij} > \boldsymbol{x}_{il}^t \boldsymbol{\beta} + \xi_{il} \quad \text{for} \quad l = 1, \ldots, k, l \neq j.$$

(By assumption this choice exists with probability one.) Then

$$p_j(\boldsymbol{X}_i, \boldsymbol{\beta}) = P(\boldsymbol{x}_{ij}^t \boldsymbol{\beta} + \xi_{ij} > \boldsymbol{x}_{il}^t \boldsymbol{\beta} + \xi_{il} \text{ for all } l \neq j,\ l = 1, \ldots, k \,|\, \boldsymbol{\beta}). \tag{1.4}$$

Here probability is in terms of the joint distribution of $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{ik})^t$, and throughout this paper we assume $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$ are independent random vectors with the same joint distribution. Although commonly the case, the components of $\boldsymbol{\xi}_i$ are not necessarily independent and identically distributed within the vector.

A Multinomial Logistic Model. If we assume that the ξ_{ij}, $j = 1, \ldots, k$, $i = 1, \ldots, n$, are i.i.d. (independent and identically distributed) from the Type I extreme value (or Gumbel) distribution with cdf (cumulative distribution function) $F(t) = \exp(-e^{-t})$, $-\infty < t < \infty$, then it can be shown that

$$p_j(\boldsymbol{X}_i, \boldsymbol{\beta}) \propto \exp(\boldsymbol{x}_{ij}^t \boldsymbol{\beta}), \quad j = 1, \ldots, k$$

(e.g., see Anderson, de Palma and Thisse (1992, p.39)). Consequently, $p_j(\boldsymbol{X}_i, \boldsymbol{\beta})$ is given by (1.2). The multinomial logistic model is one of the oldest multinomial choice models because of its closed form expression (1.3).

A Multinomial Probit Model. If we assume that the ξ_{ij} are i.i.d. $N(0,1)$ random variables, we have a multinomial probit model. In this case, the probability that person i makes choice j is

$$p_j(\boldsymbol{X}_i, \boldsymbol{\beta}) = \int_{-\infty}^{\infty} \prod_{l \neq j} \Phi(s + (\boldsymbol{x}_{ij} - \boldsymbol{x}_{il})^t \boldsymbol{\beta}) d\Phi(s),$$

where Φ is the cdf of the standard normal distribution. A special case is binary probit regression when $k = 2$, $\boldsymbol{x}_{i1} = \boldsymbol{x}_i$, $\boldsymbol{x}_{i2} = \mathbf{0}$, and ξ_{i1}, ξ_{i2} are i.i.d. $N(0, .5)$ random variables. Then $P(y_i = 1 \mid \boldsymbol{\beta}) = \Phi(\boldsymbol{x}_i^t \boldsymbol{\beta}) = 1 - P(y_i = 0 \mid \boldsymbol{\beta})$. While this property is appealing, computation can be demanding. A number of frequentist and Bayesian computational methods for these models have been proposed. See, for example, McFadden (1984), Geweke (1991), Geweke, Keane and Runkle (1994), Keane (1994), Albert and Chib (1993), McCulloch and Rossi (1994), Nobile (1998) Chib and Greenberg (1998), McCulloch, Polson and Rossi (2000) and Imai and van Dyk (2005). As far as we know, despite the computational advances, there are no general results on the existence of the MLE or propriety of the posterior for these models.

A Multinomial multivariate-t Model. There are other possible forms of the utility. For example, we might assume that the $\boldsymbol{\xi}_i$ have multivariate-t distributions with density

$$p(\boldsymbol{\xi}) = \frac{\Gamma((a+k)/2)}{\Gamma(a/2)(a\pi)^{k/2} |\boldsymbol{\Sigma}|^{1/2} [1 + a^{-1} \boldsymbol{\xi}^t \boldsymbol{\Sigma}^{-1} \boldsymbol{\xi}]^{(a+k)/2}}, \quad \boldsymbol{\xi} \in \mathbb{R}^k, \tag{1.5}$$

where $a > 0$, and $\boldsymbol{\Sigma}$ is a positive definite $k \times k$ matrix. This multivariate-t distribution has a hierarchical structure. If $\boldsymbol{\xi} \mid \delta \sim \text{MVN}_k(\mathbf{0}, \delta\boldsymbol{\Sigma})$ and δ has an inverse-gamma$(a/2, 1/2)$ distribution, namely $1/\delta \sim \chi^2_a$, the marginal density of $\boldsymbol{\xi}$ is (1.5) (cf., Anderson (1984, p.283)). Here the density of δ is given by $f_0(\delta) = \{2^{a/2}\Gamma(a/2)\}^{-1} \delta^{-a/2-1} \exp\{-1/(2\delta)\}$. When $\boldsymbol{\Sigma}$ is the $k \times k$ identity matrix $\boldsymbol{I}_k$, the probability that person i makes choice j has the form

$$p_j(\boldsymbol{X}_i, \boldsymbol{\beta}) = \int_0^{\infty} \int_{-\infty}^{\infty} \prod_{l \neq j} \Phi\Big[\frac{s + (\boldsymbol{x}_{ij} - \boldsymbol{x}_{il})^t \boldsymbol{\beta}}{\sqrt{\delta}}\Big] d\Phi(s\delta^{-\frac{1}{2}}) f_0(\delta)\, d\delta. \tag{1.6}$$

The rest of the paper is organized as follows. In Section 2, we generalize Albert and Anderson's (1984) concepts of complete separation, quasi-complete separation, and overlap for logistic classification models to general multinomial choice models. An equivalence lemma relates the overlap conditions to the conical hull of a special set of "structure vectors" associated with the problem. A "quasi-norm" constructed from the conical hull is used to prove the existence of both the MLE and the posterior in Section 3. Under mild restrictions, it is also

shown that the MLE exists if and only if the posterior with constant prior is proper. In Section 4, some conditions are given to verify overlap in choice models. Finally, Section 5 contains a short example. An online supplement contains two appendices. Proofs of lemmas on overlap are given in Appendix A, and consistency and asymptotic normality for both the MLE and the posterior are proved under some mild additional conditions in Appendix B.

2. Preliminaries

A number of authors including Haberman (1974), Wedderburn (1976), Silvapulle (1981), Silvapulle and Burridge (1986), Albert and Anderson (1984) and Santner and Duffy (1986) have considered the problem of existence of finite maximum likelihood estimates for logistic regression and classification problems. See also Amemiya (1976) and Tse (1986). Our work most closely follows Albert and Anderson (1984), and we generalize their results to multinomial choice models for the logistic case, and to more general choice models based on multivariate normal and multivariate-t choice distributions.

2.1. Separation and overlap

Albert and Anderson (1984) gave the following criteria for the logistic multinomial classification problem with $k \geq 2$. To motivate their definitions, suppose that observation i is classified to group j if

$$\boldsymbol{x}_i^t \boldsymbol{\beta}_j > \boldsymbol{x}_i^t \boldsymbol{\beta}_l \quad \text{for all } l \neq j. \tag{2.1}$$

The sample $(\boldsymbol{X}^{(n)}, \boldsymbol{y}^{(n)})$ has *complete separation* if there exist parameter vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{k-1}$ such that

$$\boldsymbol{x}_i^t \boldsymbol{\beta}_{y_i} > \boldsymbol{x}_i^t \boldsymbol{\beta}_l \quad \text{for all } l \neq y_i, i = 1, \ldots, n.$$

(By our convention, $\boldsymbol{\beta}_k = \mathbf{0}$.) In other words, complete separation means that it is possible to classify all the data points correctly by an equation of the form (2.1). The sample has *quasi-complete separation* if

$$\boldsymbol{x}_i^t \boldsymbol{\beta}_{y_i} \geq \boldsymbol{x}_i^t \boldsymbol{\beta}_l \quad \text{for all } l \neq y_i, i = 1, \ldots, n.$$

In all other cases, the sample is said to have *overlap*. In other words, there is overlap if and only if, for any choice of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{k-1}$ not all zero, there is an $\boldsymbol{x}_i$ and alternative $l \neq y_i$ such that

$$\boldsymbol{x}_i^t \boldsymbol{\beta}_{y_i} < \boldsymbol{x}_i^t \boldsymbol{\beta}_l.$$

Albert and Anderson (1984) showed that in the logistic multinomial classification problem (including logistic regression), a finite maximum likelihood estimate

exists if and only if the sample has overlap. Jacobsen (1989) derived this result from a more general theorem on discrete exponential families.

These definitions translate easily to the choice model setup given by (1.1) and (1.4). Denote the data points by $(\boldsymbol{X}^{(n)}, \boldsymbol{y}^{(n)})$, where $\boldsymbol{X}^{(n)} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is the matrix of covariates and $\boldsymbol{y}^{(n)} = (y_1, \ldots, y_n)$ is the observation vector. We will say that the sample $(\boldsymbol{X}^{(n)}, \boldsymbol{y}^{(n)})$ has *complete separation* if there is a nonzero $\boldsymbol{\beta} \in \mathbb{R}^m$ such that

$$(\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{il})^t \boldsymbol{\beta} > 0 \text{ for all } i = 1, \ldots, n, \text{ and } l \neq y_i \tag{2.2}$$

and has *quasi-complete separation* if there is a nonzero $\boldsymbol{\beta} \in \mathbb{R}^m$ such that

$$(\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{il})^t \boldsymbol{\beta} \geq 0 \text{ for all } i = 1, \ldots, n, \text{ and } l \neq y_i. \tag{2.3}$$

Otherwise, the sample $(\boldsymbol{X}^{(n)}, \boldsymbol{y}^{(n)})$ has *overlap*, i.e., for every nonzero $\boldsymbol{\beta} \in \mathbb{R}^m$, we have

$$(\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{il})^t \boldsymbol{\beta} < 0 \text{ for some } i \text{ and } l \neq y_i.$$

The intuition is the same as for the classification problem. The data have overlap if and only if there is no single rule that correctly predicts the actual choice for every observation.

We abstract the notions of separation and overlap to arbitrary sets, motivated by the following special set (called the set of *structure vectors* in Jacobsen (1989)). Define

$$\mathcal{A} = \{\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{il} : l \neq y_i, l = 1, \ldots, k, i = 1, \ldots, n\}. \tag{2.4}$$

We say that $\mathcal{A}$ has *complete separation* if there is a $\boldsymbol{\beta} \in \mathbb{R}^m$ such that

$$\boldsymbol{z}^t \boldsymbol{\beta} > 0 \text{ for all } \boldsymbol{z} \in \mathcal{A}.$$

In other words, complete separation means that $\mathcal{A}$ lies in the interior of some half-space. Similarly, $\mathcal{A}$ has *quasi-complete separation* if there is a $\boldsymbol{\beta} \in \mathbb{R}^m$ ($\boldsymbol{\beta} \neq \mathbf{0}$) with

$$\boldsymbol{\beta}^t \boldsymbol{z} \geq 0 \text{ for all } \boldsymbol{z} \in \mathcal{A}, \tag{2.5}$$

and $\mathcal{A}$ is *overlapped* if for any $\mathbf{0} \neq \boldsymbol{\beta} \in \mathbb{R}^m$, there is a $\boldsymbol{z} \in \mathcal{A}$ such that

$$\boldsymbol{z}^t \boldsymbol{\beta} < 0.$$

It turns out that the notions of separation and overlap are related to geometric concepts involving cones. We review some definitions (see Panik (1993))). A *cone* $\mathcal{C} \subset \mathbb{R}^m$ is a set of points such that if $\boldsymbol{x} \in \mathcal{C}$, then $\lambda \boldsymbol{x} \in \mathcal{C}$ for any $\lambda \geq 0$.

A cone $\mathcal{C}$ in $\mathbb{R}^m$ is called a convex cone if it is also closed under addition. For a set of vectors $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_l\}$ in $\mathbb{R}^m$, a conical (nonnegative linear) combination is

$$\boldsymbol{x} = \sum_{j=1}^{l} \lambda_j \boldsymbol{a}_j,$$

where $\lambda_j \geq 0$ for $j = 1, \ldots, l$. For a set $\mathcal{C}$ in $\mathbb{R}^m$, the *conical hull* of $\mathcal{C}$ is the collection of all conical combinations of vectors from $\mathcal{C}$, i.e.

$$coni(\mathcal{C}) = \left\{ \sum_{j=1}^{i} \lambda_j \boldsymbol{a}_j : \boldsymbol{a}_j \in C, \lambda_j \geq 0 \text{ and } i \text{ is a positive integer} \right\}.$$

Thus $coni(\mathcal{C})$ is the smallest cone containing all convex combinations of points in $\mathcal{C}$.

Example 1. If $m = 2$, $\boldsymbol{a}_1 = (1, 0)$ and $\boldsymbol{a}_2 = (0, 1)$, then $\text{coni}(\{\boldsymbol{a}_1, \boldsymbol{a}_2\}) = \{(x_1, x_2) : x_1 \geq 0, x_2 \geq 0\}$, the first quadrant. We need another vector besides $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ to form $\mathbb{R}^2$. In fact, $coni(\{\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3\}) = \mathbb{R}^2$ if and only if $\boldsymbol{a}_3 = (a_{31}, a_{32})$ with $a_{31} < 0$ and $a_{32} < 0$. In general, a set $\mathcal{C} \subset \mathbb{R}^m$ must have at least $m + 1$ elements for $coni(\mathcal{C}) = \mathbb{R}^m$.

Our final notion is a function on $\mathbb{R}^m$, associated with a set $\mathcal{C}$ when $coni(\mathcal{C}) = \mathbb{R}^m$, that has important properties of a norm. For any finite subset $\mathcal{C}$ of $\mathbb{R}^m$, define

$$\|\boldsymbol{b}\|_{\mathcal{C}} = \max_{\boldsymbol{z} \in \mathcal{C}} (-\boldsymbol{z}^t \boldsymbol{b}) = -\min_{\boldsymbol{z} \in \mathcal{C}} \boldsymbol{z}^t \boldsymbol{b}, \quad \boldsymbol{b} \in \mathbb{R}^m.$$

Clearly, the definition $\|\cdot\|_{\mathcal{C}}$ satisfies the triangle inequality, $\|\boldsymbol{b}_1 + \boldsymbol{b}_2\|_{\mathcal{C}} \leq \|\boldsymbol{b}_1\|_{\mathcal{C}} + \|\boldsymbol{b}_2\|_{\mathcal{C}}$ for any $\boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathbb{R}^m$. In order for $\|\cdot\|_{\mathcal{C}}$ to be a norm, three further conditions must hold:

$$\|\boldsymbol{b}\|_{\mathcal{C}} \geq 0, \quad \boldsymbol{b} \in \mathbb{R}^m; \tag{2.6}$$

$$\|\boldsymbol{b}\|_{\mathcal{C}} = 0 \text{ if and only if } \boldsymbol{b} = \boldsymbol{0}; \tag{2.7}$$

$$\|\alpha \boldsymbol{b}\|_{\mathcal{C}} = |\alpha| \|\boldsymbol{b}\|_{\mathcal{C}} \text{ for all } \alpha \in \mathbb{R}. \tag{2.8}$$

Property (2.8) only holds for $\alpha \geq 0$ and does not hold in general. However, when the sample has overlap, $\|\cdot\|_{\mathcal{A}}$ has the other properties of a norm.

Lemma 1. *The following conditions are equivalent.*

(i) *The sample* $(\boldsymbol{X}^{(n)}, \boldsymbol{y}^{(n)})$ *has overlap.*

(ii) $coni(\mathcal{A}) = \mathbb{R}^m$.

(iii) *Properties* (2.6) *and* (2.7) *hold.*

The proof is given in Appendix A. The lemma has a very useful consequence. Let $\|\boldsymbol{b}\| = \sqrt{b_1^2 + \cdots + b_m^2}$ denote the Euclidean norm for $\boldsymbol{b} = (b_1, \ldots, b_m)^t \in \mathbb{R}^m$. The proof of the following corollary is also given in Appendix A.

Corollary 1. *If any of the conditions of the Lemma* 1 *hold, there is a constant* $C > 0$ *such that* $C\|\boldsymbol{b}\| \leq \|\boldsymbol{b}\|_{\mathcal{A}}$ *for all* $\boldsymbol{b} = (b_1, \ldots, b_m)^t \in \mathbb{R}^m$.

2.2. Binary case

For ordinary logistic and probit regression, the classification problem with two outcomes, Silvapulle (1981) gave necessary and sufficient conditions in terms of the sample outcomes. Let

$$\mathcal{S} = coni\{\boldsymbol{x}_i : y_i = 1\} = \left\{\boldsymbol{x} : \boldsymbol{x} = \sum_{i=1, y_i=1}^{n} \lambda_i \boldsymbol{x}_i, \lambda_i \geq 0\right\},$$

$$\mathcal{F} = coni\{\boldsymbol{x}_i : y_i = 0\} = \left\{\boldsymbol{x} : \boldsymbol{x} = \sum_{i=1, y_i=0}^{n} \lambda_i \boldsymbol{x}_i, \lambda_i \geq 0\right\}.$$

Our notation differs slightly from Silvapulle's, who uses $\mathcal{S}$ and $\mathcal{F}$ to denote the relative interiors, not the cones themselves. Silvapulle's conditions for the existence of the MLE are

(S1) the rank of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is m;

(S2) either $\mathcal{S} \cap \mathcal{F} \neq \emptyset$ or one of $\mathcal{S}, \mathcal{F} = \mathbb{R}^m$,

where m is the dimension of $\boldsymbol{x}$.

Note that for the special case of binary outcomes $y_i = 0$ or 1,

$$\mathcal{A} = \{(-1)^{y_i+1}\boldsymbol{x}_i : \ i = 1, \ldots, n\}. \tag{2.9}$$

It's not hard to show the equivalence of Silvapulle's conditions to the conical hull condition in the last lemma. The proof of the next result is in Appendix A.

Lemma 2. *The conditions of Lemma* 1 *in the binary case with* $k = 2$ *are equivalent to* (S1) and (S2).

3. Existence of the MLE and the Posterior

The likelihood function of $\boldsymbol{\beta}$ based on data $(\boldsymbol{X}^{(n)}, \boldsymbol{y}^{(n)})$ is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_{y_i}(\boldsymbol{X}_i, \boldsymbol{\beta}).$$

We examine conditions for the existence of the MLE and the posterior under a constant prior. We first prove that under complete or quasi-complete separation,

the posterior is improper and the MLE does not exist. Let $G(u_1, \ldots, u_{k-1})$ be the $(k-1)$-dimensional common distribution function of $(\xi_{i1} - \xi_{ik}, \ldots, \xi_{i,k-1} - \xi_{ik})$.

Theorem 1. *Assume that G is absolutely continuous, is exchangeable in its arguments, and has a density that is positive on $\mathbb{R}^{m-1}$. Then the MLE exists and is finite if and only if there is overlap in the sample $(\boldsymbol{X}^{(n)}, \boldsymbol{y}^{(n)})$.*

Proof. We first consider necessity and argue by contradiction. If there is no overlap, we have quasi-complete (which includes complete) separation, i.e. there is a nonzero $\boldsymbol{\beta}_* \in \mathbb{R}^m$ so that

$$(\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{il})^t \boldsymbol{\beta}_* \geq 0 \text{ for } i = 1, \ldots, n, \text{ and } l \neq y_i. \tag{3.1}$$

If the rank of $\mathcal{A}$ is m, there must be an observation i_* and an alternative $l_* \neq y_{i_*}$ such that $(\boldsymbol{x}_{i_* y_{i_*}} - \boldsymbol{x}_{i_* l_*})^t \boldsymbol{\beta}_* > 0$. For any finite $\boldsymbol{\beta} \in \mathbb{R}^m$,

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} G((\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{i1})^t \boldsymbol{\beta}, \ldots, (\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{ik})^t \boldsymbol{\beta}) \\ &< \prod_{i=1}^{n} G((\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{i1})^t (\boldsymbol{\beta} + \boldsymbol{\beta}_*), \ldots, (\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{ik})^t (\boldsymbol{\beta} + \boldsymbol{\beta}_*)) \\ &= L(\boldsymbol{\beta} + \boldsymbol{\beta}_*), \end{aligned}$$

where the inequality holds because (3.1) holds, strict inequality holds for some i_* and $l_* \neq y_{i_*}$, and G is strictly increasing in its coordinates since the density is assumed positive. Thus no finite $\boldsymbol{\beta}$ can maximize $L(\boldsymbol{\beta})$. On the other hand, if equality holds in (3.1) for all i and $l \neq y_i$, then $\mathcal{A}$ must have rank less than k. In this case, $L(\boldsymbol{\beta}) = L(\boldsymbol{\beta} + t\boldsymbol{\beta}_*)$ for all $\boldsymbol{\beta} \in \mathbb{R}^m$ and $t \in \mathbb{R}$, again contradicting the assumption that the MLE must be finite.

To prove sufficiency, note that for any fixed $\boldsymbol{\beta}$ and any $j = 1, \ldots, k$,

$$\begin{aligned} p_j(\boldsymbol{X}_i, \boldsymbol{\beta}) &= P(\xi_{il} - \xi_{ij} \leq (\boldsymbol{x}_{ij} - \boldsymbol{x}_{il})^t \boldsymbol{\beta}, \text{ for all } l \neq j \mid \boldsymbol{\beta}) \\ &\leq P(\xi_{is} - \xi_{ij} \leq (\boldsymbol{x}_{ij} - \boldsymbol{x}_{is})^t \boldsymbol{\beta} \mid \boldsymbol{\beta}) \end{aligned}$$

for any $s \neq j$. Define

$$H(t) = P(\min_{j \neq l = 1, \ldots, k} (\xi_{1l} - \xi_{1j}) \leq t), \quad t \in \mathbb{R}.$$

Then

$$p_j(\boldsymbol{X}_i, \boldsymbol{\beta}) \leq \min_{1 \leq l \leq k,\ l \neq y_i} H((\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{il})^t \boldsymbol{\beta})$$

and, since H is bounded by 1,

$$\begin{aligned} L(\boldsymbol{\beta}) &\leq \min_{1 \leq l \leq k,\ l \neq y_i, 1 \leq i \leq n} H((\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{il})^t \boldsymbol{\beta}) \\ &= H(\min_{1 \leq l \leq k,\ l \neq y_i,\ 1 \leq i \leq n} (\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{il})^t \boldsymbol{\beta}), \end{aligned}$$

where the second equality holds because H is a monotone function. From the assumption and Corollary 1, we know that there is a constant $C > 0$ such that

$$L(\boldsymbol{\beta}) \leq H(-C\|\boldsymbol{\beta}\|), \tag{3.2}$$

where $\|\boldsymbol{\beta}\|$ is the Euclidean norm of $\boldsymbol{\beta}$. Fix an arbitrary vector $\boldsymbol{\beta}^* \in \mathbb{R}^m$. Then there is a constant $M > 0$ so that for any $\|\boldsymbol{\beta}\| \geq M$,

$$L(\boldsymbol{\beta}) \leq H(-C\|\boldsymbol{\beta}\|) \leq H(-CM) < L(\boldsymbol{\beta}^*).$$

Therefore

$$\sup_{\boldsymbol{\beta}\in\mathbb{R}^m} L(\boldsymbol{\beta}) = \sup_{\|\boldsymbol{\beta}\|\leq M} L(\boldsymbol{\beta}).$$

Since $L(\boldsymbol{\beta})$ is continuous, its maximum exists and any MLE is finite.

Remark. For the special case of the logistic multinomial choice model (1.2), the theorem follows directly from Jacobsen's (1989) results on discrete exponential families. Jacobsen's theorem also shows uniqueness of the MLE when it exists for the logistic case.

To illustrate the meaning of separation, quasi-separation and overlap, consider the following examples.

Example 2. Consider ordinary logistic regression with two observations, both at the same $x > 0$, with one success and one failure. Then

$$L(\boldsymbol{\beta}) = \left\{\frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}\right\}\left\{\frac{1}{1 + \exp(\beta_0 + \beta_1 x)}\right\},$$

whose maximum of 0.25 is achieved for any $\beta_0 + \beta_1 x = 0$. From (2.9), $\mathcal{A} = \{(1, x), (-1, -x)\}$, which is quasi-complete. In particular, $\boldsymbol{z}^t\boldsymbol{\beta} = 0$ for all $\boldsymbol{z} \in \mathcal{A}$ if $\boldsymbol{\beta} = (-x, 1)$. Hence $L(\boldsymbol{\beta})$ achieves its maximum, but the set on which the maximum is attained is unbounded.

Example 3. Suppose a third observation is added to the data set, a success at $2x$. Now $\mathcal{A} = \{(1, x), (-1, -x), (1, 2x)\}$, which is still quasi-completely separated (using the same $\boldsymbol{\beta}$) but $\mathcal{A}$ has full rank. For this case,

$$L(\boldsymbol{\beta}) = \frac{\exp(\beta_0 + \beta_1 x)\exp(\beta_0 + 2\beta_1 x)}{[1 + \exp(\beta_0 + \beta_1 x)]^2[1 + \exp(\beta_0 + 2\beta_1 x)]}.$$

It's easy to see that $L(\boldsymbol{\beta}) < 0.25$ for all $\boldsymbol{\beta}$. However, $L(\boldsymbol{\beta}_s) \to 0.25$ as $s \to \infty$ if $\boldsymbol{\beta}_s = (-sx, s)$. No finite MLE exists in this case.

Example 4. Finally, suppose there are again three observations but with one success at x and failures at 0 and $2x$. Then $\mathcal{A} = \{(1, x), (-1, 0), (-1, -2x)\}$ and $coni(\mathcal{A}) = \mathbb{R}^2$. In this case, the likelihood

$$L(\boldsymbol{\beta}) = \frac{\exp(\beta_0 + \beta_1 x)}{[1 + \exp(\beta_0)][1 + \exp(\beta_0 + \beta_1 x)][1 + \exp(\beta_0 + 2\beta_1 x)]}.$$

is maximized with $\beta_1 = 0$ and $1/\{1 + \exp(\beta_0)\} = 2/3$.

Example 2 illustrates that if $\mathcal{A}$ does not have full rank, the set of MLE's is not bounded. In Example 3, $\mathcal{A}$ does have full rank, but $coni(\mathcal{A})$ is not the whole parameter space, so the MLE does not exist. In Example 4, $\mathcal{A}$ has full rank, $coni(\mathcal{A}) = \mathbb{R}^m$, and the MLE exists and is finite.

Our next theorem gives conditions for the existence of a proper posterior. For simplicity, we again assume that the distribution of $\boldsymbol{\xi}_i$ is exchangeable. Our sufficient condition assumes only a moment condition on the components ξ_{ij}.

Theorem 2. *Assume that a constant prior for $\boldsymbol{\beta}$ is used and G is exchangeable in its arguments.* (a) *If there is quasi-complete separation in the sample $(\boldsymbol{X}^{(n)}, \boldsymbol{y}^{(n)})$ and the cdf G is continuous, the posterior is improper.* (b) *If there is overlap in the sample $(\boldsymbol{X}^{(n)}, \boldsymbol{y}^{(n)})$ and $E(|\xi_{1j}|^m) < \infty$ for $j = 1, \ldots, k$, the posterior of $\boldsymbol{\beta}$ is proper.*

Proof. To prove (a), suppose that there is a $\boldsymbol{\beta}_* \in \mathbb{R}^m$ satisfying (3.1). Without loss of generality, assume that $\|\boldsymbol{\beta}_*\| = 1$. Choose $\boldsymbol{\alpha}_* \in \mathbb{R}^m$ such that $L(\boldsymbol{\alpha}_*) > 0$. Since $L(\boldsymbol{\alpha})$ is continuous, there exists a constant $C > 0$ and a bounded neighborhood of $\boldsymbol{\alpha}_*$, say U, such that $L(\boldsymbol{\alpha}) > C$ for any $\boldsymbol{\alpha} \in U$. Since U is bounded, there is a constant M so that $\|\boldsymbol{\alpha}\| \leq M$ for any $\boldsymbol{\alpha} \in U$. Define $B = \{\boldsymbol{\alpha} + t\boldsymbol{\beta}_* : \boldsymbol{\alpha} \in U, t > 0\}$. Then for any $\boldsymbol{\beta} = \boldsymbol{\alpha} + t\boldsymbol{\beta}_* \in B$, and for any $l \neq y_i, i = 1, \ldots, n$,

$$\begin{aligned}(\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{il})^t\boldsymbol{\beta} &= (\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{il})^t(\boldsymbol{\alpha} + t\boldsymbol{\beta}_*) \\ &= (\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{il})^t\boldsymbol{\alpha} + t(\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{il})^t\boldsymbol{\beta}_* \\ &\geq (\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{il})^t\boldsymbol{\alpha},\end{aligned}$$

where the last inequality follows from (3.1).

$$\begin{aligned}L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} G((\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{i1})^t\boldsymbol{\beta}, \ldots, (\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{ik})^t\boldsymbol{\beta}) \\ &\geq \prod_{i=1}^{n} G((\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{i1})^t\boldsymbol{\alpha}, \ldots, (\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{ik})^t\boldsymbol{\alpha}) \\ &= L(\boldsymbol{\alpha}) = C\end{aligned}$$

for all $\boldsymbol{\beta} \in B$. Clearly the Lebesgue measure $\lambda(B)$ is infinite, hence

$$\int_{\mathbb{R}^m} L(\boldsymbol{\beta})\, d\boldsymbol{\beta} \;\geq\; \int_B L(\boldsymbol{\beta})\, d\boldsymbol{\beta} \;\geq\; C\lambda(B) = \infty.$$

This proves (a).

For part (b), it follows from the sufficiency proof of Theorem 1 that there is a constant $C > 0$ such that (3.2) holds, and

$$\int_{\mathbb{R}^m} L(\boldsymbol{\beta})\, d\boldsymbol{\beta} \leq \int_{\mathbb{R}^m} H(-C\|\boldsymbol{\beta}\|)\, d\boldsymbol{\beta}.$$

By using the polar transformation, we get

$$\begin{aligned}
\int_{\mathbb{R}^m} L(\boldsymbol{\beta}) d\boldsymbol{\beta} &\leq C_1 \int_0^\infty r^{m-1} H(-Cr)\, dr \\
&= C_2 \int_0^\infty r^{m-1} H(-r)\, dr \\
&= C_2 \int_0^\infty r^{m-1} \int_{-\infty}^{-r} dH(s)\, dr \\
&= C_2 \int_{-\infty}^0 \left\{ \int_0^{-s} r^{m-1} dr \right\} dH(s) \\
&\leq C_3 \int_{-\infty}^\infty |s|^m\, dH(s) \\
&\leq C_3 E \left| \min_{j \neq l = 1,\ldots,k} (\xi_{1l} - \xi_{1j}) \right|^m,
\end{aligned}$$

where the C_i are finite positive constants. Since ξ_{1j} has finite mth moment, the right hand side is finite, proving (b).

Remark 2. Although the ξ_{ij}, $j = 1, \ldots, k$, are often i.i.d. random variables, they need not be independent or identically distributed for the conditions of the theorem to hold.

Note that the ξ_{ij} are i.i.d. with the standard extreme-value distribution for the logistic model and the normal distribution for the probit model. Since both distributions have all moments, the next result follows from Theorems 1 and 2 immediately.

Theorem 3. *For the multinomial logistic or probit choice model, the following conditions are equivalent.*

(i) *There is overlap in the sample* $(\boldsymbol{X}^{(n)}, \boldsymbol{y}^{(n)})$.

(ii) *The MLE of* $\boldsymbol{\beta}$ *exists and is finite.*

(iii) *The posterior of* $\boldsymbol{\beta}$ *is proper under the constant prior.*

Raghavan and Cox (1998) showed that (ii) and (iii) are equivalent for a binary logistic model, which is a special case here.

Remark 3. We should point out that the moment condition of the theorem is sufficient and convenient, but not necessary. Since our main concerns are the multinomial logistic and probit models, the moment condition suffices here. However, for the multinomial multivariate-t model given by (1.1) and (1.6), $E(|\xi_{ij}|^m)$ is finite if and only if $a - m > 0$. A weaker condition may be possible; we do not explore that here.

Note that the constant prior is also not necessary here. In practice, one might use a proper prior such as a normal distribution or a partially informative prior such as the one in Sun, Tsutakawa and Speckman (1999), defined by

$$p(\boldsymbol{\beta}) \propto \exp(-\frac{1}{2\delta}\boldsymbol{\beta}^t \boldsymbol{B}\boldsymbol{\beta}),$$

where $\delta > 0$ and $\boldsymbol{B}$ is a nonnegative definite symmetric matrix. Such a prior is bounded above by a constant. The proof of Theorem 2(b) easily extends to bounded priors, and we have the following.

Corollary 2. *Assume that there is overlap in the sample* $(\boldsymbol{X}^{(n)}, \boldsymbol{y}^{(n)})$ *and* $E(|\xi_{1j}|^m) < \infty$ *for* $j = 1, \ldots, k$. *If the prior of* $\boldsymbol{\beta}$ *is bounded, the posterior of* $\boldsymbol{\beta}$ *is proper.*

4. Verifying Overlap in Choice Models

From a practical standpoint, it may not be clear if a data set has overlap so the MLE exists. A number of authors have addressed this problem in the context of logistic regression and the multinomial logistic model (1.3). Albert and Anderson (1984) give a detailed discussion of various strategies. Translated to our setting, they note that the problem of determining complete separation (2.2) or quasi-complete separation (2.3) is a linear programming problem. Further details on implementing the linear programming problem are given by Silvapulle and Burridge (1986). Albert and Anderson (1984) also proposed methods for monitoring maximum likelihood iterations to determine if complete separation exists. When complete separation exists, the logistic discriminant rule will eventually give perfect discrimination. Monitoring the number of correct classifications during iteration can detect this case. The situation with quasi-complete separation is more complicated, but Albert and Anderson (1984) suggest a more sophisticated monitoring scheme that appears effective in detecting problem cases. Lesaffre and Albert (1989) proposed detecting problem data sets based on monitoring the number of correctly classified cases and the standard errors of the estimated parameters from standard output of maximum likelihood programs.

These results clearly carry over to the logistic multinomial choice model (1.2). Moreover, the problem of detecting complete or quasi-complete separation is independent of the choice of latent variable distribution, so the Lesaffre and Albert (1989) method based on logistic regression could be used for detecting problems for probit or Bayesian models, for example.

In some cases, it might be possible to easily verify overlap based on certain simple rules. For example, consider a subset of the sample, say $I \subset \{1, \ldots, n\}$. Let $\mathcal{A}_I$ denote the set of structure vectors for this restricted sample, i.e.,

$$\mathcal{A}_I = \{\boldsymbol{x}_{iy_i} - \boldsymbol{x}_{il} : l \neq y_i, l = 1, \ldots, k, i \in I\}.$$

Clearly, if $\mathcal{A}_I$ has overlap, then $\mathcal{A}$ must as well. In describing applications, we distinguish two cases. Let Case I denote models incorporating a separate constant term for all but one possible choice. Thus the explanatory vectors $\boldsymbol{x}_{ij}$ might have the form

$$\begin{aligned}
\boldsymbol{x}_{i1}^t &= (1, 0, \ldots, 0, \boldsymbol{z}_{i1}^t), \\
\boldsymbol{x}_{i2}^t &= (0, 1, \ldots, 0, \boldsymbol{z}_{i2}^t), \\
\boldsymbol{x}_{i,p-1}^t &= (0, \ldots, 0, 1, \boldsymbol{z}_{i,p-1}^t) \\
\boldsymbol{x}_{ip}^t &= (0, 0, \ldots, 0, \boldsymbol{z}_{ip}^t),
\end{aligned}$$

where the $\boldsymbol{z}_{ip}$ are covariate vecors. Let $\boldsymbol{\beta}^t = (\beta_1, \beta_2, \ldots, \beta_{p-1}, \boldsymbol{\gamma}^t)$, so that

$$\begin{aligned}
p_j(\boldsymbol{X}_i, \boldsymbol{\beta}) &\propto \exp(\beta_j + \boldsymbol{\gamma}^t \boldsymbol{z}_{ij}), 1 \leq j < p, \\
p_p(\boldsymbol{X}_i, \boldsymbol{\beta}) &\propto \exp(\boldsymbol{\gamma}^t \boldsymbol{z}_{ip}).
\end{aligned}$$

Case II denotes a model without constant parameters for each choice. In some applications, such as the transportation study in Appendix B, models are constructed without constant terms because one wants to estimate the probability of future choices based solely on attributes of the choice itself and possible covariates dependent on interactions between the ith case and the choices. In other cases, it makes sense to include the constant terms.

In Case II, with no constant terms, it is possible that there is overlap in a subset of the sample belonging to as few as two possible choices. Choosing the two most frequent outcomes, for example, one could run a simple simple binary logistic regression for the restricted subset. Following the rules of Lesaffre and Albert (1989), if the algorithm converges and the standard errors are satisfactory, one could conclude that there is overlap in the binary case, which automatically implies overlap in the complete data set.

For Case I, we have the following result. Let $H_j = \{i : y_i = j\}, j = 1, \ldots, p$, i.e., H_j is the portion of the sample with choice j.

Theorem 4. *Suppose the following two conditions hold:*

(i) $H_j \neq \emptyset$, $j = 1, \ldots, p$;

(ii) *there exists a pair of choices $j \neq k$ such that the subset of the sample $H_j \cup H_k$ has overlap with respect to the binary choice problem.*

Then the entire sample has overlap.

Proof. Assume that $j, k < p$. (The case $j = p$ or $k = p$ is analogous and left to the reader.) Without loss of generality, let $j = 1$ and $k = 2$. We argue by contradiction. Suppose $\mathcal{A}$ has quasi-complete separation, i.e., there exists a nonzero vector $\boldsymbol{b}^t = (b_1, \ldots, b_{p-1}, \boldsymbol{c}^t)$ such that $\boldsymbol{b}^t \boldsymbol{v}_{ij} \geq 0$ for all $\boldsymbol{v}_{ij} \in \mathcal{A}$. The restricted set of structure vectors $\mathcal{A}_{H_1 \cup H_2}$ consists exactly of vectors of the form

$$\begin{aligned} \boldsymbol{v}_{i1} &= \boldsymbol{x}_{i1} - \boldsymbol{x}_{i2} = (1, -1, 0, \ldots, 0, \boldsymbol{z}_{i1}^t - \boldsymbol{z}_{i2}^t)^t, i \in H_1, \\ \boldsymbol{v}_{i2} &= \boldsymbol{x}_{i2} - \boldsymbol{x}_{i1} = (-1, 1, 0, \ldots, 0, \boldsymbol{z}_{i2}^t - \boldsymbol{z}_{i1}^t)^t, i \in H_2. \end{aligned}$$

This restricted set must also have quasi-complete separation, so

$$\begin{aligned} &b_1 - b_2 + \boldsymbol{c}^t(z_{i1} - z_{i2}) \geq 0 \text{ for all } i \in H_1, \\ &b_2 - b_1 + \boldsymbol{c}^t(z_{i2} - z_{i1}) \geq 0 \text{ for all } i \in H_2. \end{aligned} \tag{4.1}$$

On the other hand, by assumption, the binary problem with structure vectors $\tilde{\boldsymbol{v}}_{i1} = (1, \boldsymbol{z}_{i1}^t - \boldsymbol{z}_{i2}^t)^t$ for $i \in H_1$ and $\tilde{\boldsymbol{v}}_{i2} = (-1, \boldsymbol{z}_{i2}^t - \boldsymbol{z}_{i1}^t)^t$ for $i \in H_2$ has overlap. Thus if $\tilde{\boldsymbol{b}}^t \tilde{\boldsymbol{v}}_{ij} \geq 0$ for all $\tilde{\boldsymbol{v}}_{ij}$, then $\tilde{\boldsymbol{b}} = 0$. From (4.1), we conclude $b_1 = b_2$ and $\boldsymbol{c} = 0$. Next, consider choices $j < p$ and $k = p$. By assumption, there is at least one observation for each choice, so without loss of generality, let $j = 1$. Then $\mathcal{A}$ contains at least one vector of the form $\boldsymbol{v}_{ip} = (1, 0, \ldots, 0, \boldsymbol{z}_{i1}^t - \boldsymbol{z}_{ip}^t)^t$ for some $i \in H_1$, and at least one of the form $\boldsymbol{v}_{\ell 1} = (-1, 0, \ldots, 0, \boldsymbol{z}_{\ell p}^t - \boldsymbol{z}_{\ell 1}^t)^t$ for some $\ell \in H_p$. But we have already shown that $\boldsymbol{c} = 0$, so $\boldsymbol{b}^t \boldsymbol{v}_{ip} = b_1 \geq 0$ and $\boldsymbol{b}^t \boldsymbol{v}_{\ell 1} = -b_1 \geq 0$, implying $b_1 = 0$. Similarly, we conclude $b_2 = \cdots = b_{p-1} = 0$, contradicting the existence of nonzero $\boldsymbol{b}$. This shows that the complete sample must have overlap.

5. Application

We have computed the MLE and a Bayesian estimate for a simplified choice model based on data from the 1994/95 Portland Activity/Travel Survey. The Portland, Oregon metropolitan region is divided into 1244 travel activity zones. We chose $k = 100$ zones near downtown as possible work locations. A total of $n = 1386$ individuals from the sample worked in one of these zones. We used two factors to predict work zone choice, the total number of employees in zone j and the distance from home to the center of zone j. Thus the probability that

Table 1. The MLE and the Posterior Quantities of β_j for a Multinomial Logit Model

	MLE	Post. Mean	Post Variance	Post. Stan. Dev.
β_1	0.98637	0.98682	0.00057	0.0239
β_2	−2.55467	−2.54562	0.03428	0.1851

person i works in zone j is a function of $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2})^t$, where for $i = 1, \ldots, n$, $j = 1, \ldots, k$,

$$x_{ij1} = t_j = \log(\text{number employees at zone } j),$$
$$x_{ij2} = z_{ij} = \text{distance (in km.) from the } i\text{th person's home to zone } j.$$

Note that x_{ij1} depends only on j, while $x_{ij2} = z_{ij}$ depends on both i and j.

To verify overlap in this sample, we used the method of Case II in Section 4. The set of structure vectors $\mathcal{A}$ for the first two locations showed obvious overlap, hence there is overlap for the full data set. One can also observe that the MLE iterations converge with finite asymptotic variances as shown in the table below, again indicating overlap according to the criteria of Albert and Anderson (1984) and Lesaffre and Albert (1989).

We fit the multinomial logistic model

$$p_j(\boldsymbol{X}_i, \boldsymbol{\beta}) = \frac{\exp(\beta_1 t_j + \beta_2 z_{ij})}{\sum_{\ell=1}^{k} \exp(\beta_1 t_\ell + \beta_2 z_{i\ell})}$$

to the probability that person i works in zone j. Given observations $(y_1, \ldots, y_n)$, the likelihood function of $\boldsymbol{\beta} = (\beta_1, \beta_2)$ is

$$L(\boldsymbol{\beta}) = p(\boldsymbol{y} \mid \boldsymbol{\beta}) = \prod_{i=1}^{n} p_{y_i}(\boldsymbol{X}_i, \boldsymbol{\beta}) \;=\; \frac{\exp(\beta_1 \sum_{i=1}^{n} t_{y_i} + \beta_2 \sum_{i=1}^{n} z_{iy_i})}{\prod_{i=1}^{n} \sum_{j=1}^{k} \exp(\beta_1 t_j + \beta_2 z_{ij})}.$$

The MLE is given in Table 1.

To compute Bayesian estimates, we used a constant prior on $\boldsymbol{\beta}$ and Gibbs sampling. We have the following full conditional distributions:

$$f_1(\beta_1 \mid \beta_2; \text{data}) \propto \frac{\exp(\beta_1 \sum_{i=1}^{n} t_{y_i})}{\prod_{i=1}^{n} \sum_{j=1}^{k} \exp(\beta_1 \sum_{i=1}^{n} t_j + \beta_2 z_{ij})},$$
$$f_2(\beta_2 \mid \beta_1; \text{data}) \propto \frac{\exp(\beta_2 \sum_{i=1}^{n} z_{iy_i})}{\prod_{i=1}^{n} \sum_{j=1}^{k} \exp(\beta_1 \sum_{i=1}^{n} t_j + \beta_2 z_{ij})}.$$

These conditional densities are all log-concave, so the adaptive algorithm from Gilks and Wild (1992) can be used. We used 5,000 samples for burn-in and obtained another 10,000 samples to estimate the posterior. Estimates of the

posterior means, variances and standard deviations of β_1 and β_2 are given in Table 1. In this example, the MLE and Bayesian estimates of β_j are quite close.

Acknowledgement

This research was supported in part by the US Federal Highway Administration, under subcontract E77690017-3Y from Los Alamos National Laboratory to the National Institute of Statistical Sciences, by NSF grants SES-0351523 and SES-0720229, by NIH grant R01-MH071418, and by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2005-070-C00021).

The authors are grateful to an associate editor and the referees, especially to one referee for the reference to Jacobsen (1989), and for comments that prompted Section 4.

References

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**, 1-10.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88**, 669-679.

Amemiya, T. (1976). The maximum likelihood, the minimum chi-square and the nonlinear weighted least-squares estimator in the general qualitative response model. *J. Amer. Statist. Assoc.* **71**, 347-351.

Anas, A. (1983). Discrete choice theory, information theory and the multinomial logit and gravity models. *Transportation Research B* **17**, 13-23.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. John Wiley, New York.

Anderson, S. P., de Palma, A. and Thisse, J. F. (1992). *Discrete Choice Theory of Product Differentiation*. The MIT Press, Cambridge, Massachusetts.

Ben-Akiva, M. and Lerman, S. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. The MIT Press, Cambridge, Massachusetts.

Chen, M.-H., Ibrahim, J. G. and Shao, Q.-M. (2004). Propriety of the posterior distribution and existence of the MLE for regression models with covariates missing at random. *J. Amer. Statist. Assoc.* **99**, 421-438.

Chib, S. and Greenberg, E. (1998). Analysis of multinomial probit models. *Biometrika* **85**, 347-361.

Geweke, J. (1991). Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints. *Computer Science and Statistics: Proceedings of the* 23*rd Symposium on the Interface*, 571-578.

Geweke, J., Keane, M. and Runkle, D. (1994). Alternative computational approaches to inference in the multinomial probit model. *Rev. Econom. Statist.* **76**, 609-632.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* **41**, 337-348.

Haberman, S. J. (1974). Log-linear models for frequency tables derived by indirect observation: Maximum likelihood equations. *Ann. Statist.* **74**, 911-924.

Hobert, J. P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *J. Amer. Statist. Assoc.* **91**, 1461-1473.

Imai, K. and van Dyk, D. A. (2005). A Bayesian analysis of the multinomial probit model using marginal data augmentation. *J. Econometrics* **124**, 311-334.

Jacobsen, M. (1989). Existence and unicity of MLEs in discrete exponential family distributions. *Scand. J. Statist.* **16**, 335-349.

Keane, M. P. (1994). A computationally practical simulation estimator for Panel data. *Econometrica* **62**, 95-116.

Lesaffre, E. and Albert, A. (1989). Partial separation in logistic discrimination. *J. Roy. Statist. Soc. Ser. B* **51**, 109-116.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. John Wiley, New York.

McCulloch, R. E., Polson, N. G. and Rossi, P. E. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics* **99**, 173-193.

McCulloch, R. E. and Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *J. Econometrics* **64**, 207-240.

McFadden, D. (1984). Econometric analysis of qualitative response models. In *Handbook of Econometrics, Vol II* (Edited by Z. Griliches and M. Intriligator), 1395-1457. North-Holland, New York.

Nobile, A. (1998). A hybrid Markov chain for the Bayesian analysis of the multinomial probit model. *Statist. Comput.* **8**, 229-242.

Panik, M. J. (1993). *Fundamentals of Convex Analysis, Duality, Separation Representation, and Resolution*. Kluwer Academic Publishers, Boston.

Raghavan, N. and Cox, D. D. (1998). Analysis of the posterior for spline estimators in logistic regression. *J. Statist. Plann. Inference* **71**, 117-136.

Santner, T. J. and 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**, 755-758.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *J. Roy. Statist. Soc. Ser. B* **43**, 310-313.

Silvapulle, M. J. and Burridge, J. (1986). Existence of maximum likelihood estimates in regression models for grouped and ungrouped data. *J. Roy. Statist. Soc. Ser. B* **48**, 100-106.

Sun, D., Tsutakawa, R. K. and Speckman, P. L. (1999). Posterior distribution of hierarchical models using CAR(1) distributions. *Biometrika* **86**, 341-350.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review* **34**, 278-286.

Tse, S.-K. (1986). On the existence and uniqueness of maximum likelihood estimates in polytomous response models. *J. Statist. Plann. Inference* **14**, 269-273.

Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63**, 27-32.

Department of Statistics, University of Missouri-Columbia, Columbia, Missouri 65211, U.S.A.
E-mail: speckmanp@missouri.edu

Department of Statistics, Seoul National University, Seoul, Korea.
E-mail: leejyc@gmail.com

Department of Statistics, University of Missouri-Columbia, Columbia, Missouri 65211, U.S.A.
E-mail: sund@missouri.edu

(Received November 2006; accepted October 2007)