MAXIMUM LIKELIHOOD ESTIMATION OF A BINARY CHOICE MODEL WITH RANDOM COEFFICIENTS OF UNKNOWN DISTRIBUTION

by

Hidehiko Ichimura and T. Scott Thompson

Discussion Paper No. 268, April 1993

Center for Economic Research
Department of Economics
University of Minnesota
Minneapolis, MN 55455

Maximum Likelihood Estimation of a Binary Choice Model with Random Coefficients of Unknown Distribution

Hidehiko Ichimura and T. Scott Thompson

Department of Economics
University of Minnesota

March, 1993

Mailing Address: 1035 Management & Economics
271 19th Avenue South
Minneapolis, MN 55455

Telephone: (612) 625-4512 (Ichimura)
(612) 625-0119 (Thompson)

e-mail: ichimura@atlas.socsci.umn.edu
thompson@atlas.socsci.umn.edu

Abstract

We consider a binary response model $y_i = 1\{ x_i'\beta_i + \varepsilon_i \geq 0 \}$ with x_i independent of the unobservables (β_i, ε_i). No finite-dimensional parametric restrictions are imposed on F_0, the joint distribution of (β_i, ε_i). A nonparametric maximum likelihood estimator for F_0 is shown to be consistent. We analyze some conditions under which F_0 is or is not identified. We find that certain moments of F_0 are not identified, even when the model is normalized by fixing one variance. The correlation matrix of (β_i, ε_i) is not identified. We also provide some Monte Carlo evidence on the small sample performance of our estimator.

1. INTRODUCTION

The standard econometric model of binary choice postulates that for each individual i (chosen randomly from a large population) an observed choice variable y_i is related to a K×1 vector of observables x_i by the equation

$$y_i = 1\{ x_i'\beta_i + \varepsilon_i \geq 0 \} \tag{1}$$

where β_i (a K×1 vector) and ε_i are unobserved. Here $1\{\cdot\}$ denotes the zero-one binary indicator function. This model is sometimes motivated by assuming that individual i receives an indirect utility U_{ij} when alternative j is chosen, and therefore chooses $y_i = 1$ if and only if $U_{i1} \geq U_{i0}$. If there are versions of indirect utility that satisfy the separable specification

$$U_{ij} = z_{ij}'\beta_i + \nu_{ij} \tag{2}$$

then we obtain (1) by setting $x_i = z_{i1} - z_{i0}$ and $\varepsilon_i = \nu_{i1} - \nu_{i0}$.

In this context β_i is a vector of the marginal utilities associated with the observed variables x_i (e.g. prices), and ε_i collects all other unobserved determinants of indirect utility.

Until quite recently, computational restrictions prevented econometricians from modeling heterogeneity in the population except through the random intercept term ε_i. Typically it has been assumed that $\beta_i = \bar{\beta}$ for all i, where $\bar{\beta}$ is a fixed parameter vector to be estimated. More recently many of the computational problems have been overcome. For example, McFadden (1989) proposed a simulation estimator for a version of the model in which the β_i terms are also random. Earlier, Albright, Lerman and Manski (1977), and Hausman and Wise (1978) proposed other estimation methods for similar models. However, all of these authors suggested methods that require the

econometrician to specify the joint distribution of β_i and ε_i up to a finite number of parameters.

In this paper we consider a model that allows for heterogeneity in the slope parameters β_i as well as the intercept term ε_i, and yet does not require a parametric specification for the joint distribution of these unobserved terms. We assume that x_i, β_i and ε_i are all random and that β_i and ε_i are jointly independent of x_i. We propose a nonparametric maximum likelihood estimator for the distribution function F_0 of β_i under the assumption that F_0 is an element of $\mathcal{F}$, an arbitrary space of distribution functions. We analyze some conditions under which F_0 is or is not identified relative to $\mathcal{F}$. We find that certain moments of F_0 are not identified when $\mathcal{F}$ is large enough. This remains true even when β_i is normalized by fixing one variance. The correlation matrix of β_i is not identified. Our Theorem 3 provides some simple, sufficient conditions for identification.

Given identification, we prove consistency of the maximum likelihood estimator. We also provide some Monte Carlo evidence on the small sample performance of our estimator.

In the remainder of this introduction we provide a brief motivation for our model and relate it to other literature on binary choice. Section 2 considers identification issues for our model. Section 3 introduces the maximum likelihood estimator and discuss its geometric structure under our identifying conditions. Section 4 establishes consistency of the estimator. Section 5 presents some example calculations of the maximum likelihood estimator and reports the results of some Monte Carlo experiments. We conclude with some interpretation of our results and some suggestions for further research.

Prior restrictions on binary choice models.

It is well-known that imposition of incorrect restrictions on the joint distribution of x_i, β_i and ε_i in model (1) typically leads to inconsistent estimates of parameters of interest. However, economic theory rarely (if ever) provides guidance about the form of this joint distribution. On the other hand, in the absence of a priori restrictions, the joint distribution of y_i and x_i is also unrestricted. In this case (1) is vacuous, and the simple interpretations that make it appealing are not available.

Restrictions on the joint distributions in (1) are also useful when the research objective is prediction of y_i conditional on x_i. For example, the assumption that the unobservables are independent of x_i places restrictions on the conditional probabilities that $y_i = 1$. The restrictions will sometimes permit the probabilities to be estimated more efficiently than is possible when one takes an unrestricted nonparametric approach to this problem.

Thus applied econometricians have sought to find restrictions on the joint distribution of x_i and unobservables that are simultaneously (i) as weak as possible (in order to avoid introducing specification errors) and (ii) strong enough to permit interesting inferences about the choice model consistent with (1).

Most often it is assumed that $\beta_i = \bar{\beta}$, a fixed parameter to be estimated, and that ε_i is random with a known distribution, independent of x_i. Typically the standard normal or logistic distributions are assumed, leading to the classical probit and logit models respectively. Estimation of these models is straightforward using maximum-likelihood methods.

Several extensions of the classical models have received attention. Random coefficient versions of the model assume that both β_i and ε_i are random and jointly independent of x_i. Compared to a model with β_i fixed, this model is consistent with a greater range of unobserved heterogeneity in the population from which the sample is drawn. We are aware only of examples in which the distribution of $(\beta_i', \varepsilon_i)'$ is multivariate normal and ε_i is uncorrelated with the elements of β_i. See for example Albright, Lerman and Manski (1977), or Hausman and Wise (1978).[1] In this context it is natural to rewrite (1) as

$$y_i = 1\{ x_i'\bar{\beta} + u_i \geq 0 \} \tag{3}$$

where $\bar{\beta} = E(\beta_i)$ is the parameter of main interest and $u_i = x_i'(\beta_i - \bar{\beta}) + \varepsilon_i$ is a heteroskedastic composite error term. The parameter $\Sigma = Var(\beta_i)$ also must be estimated in order to make (conditional) predictions about y_i. The computational expense of computing orthant probabilities of the multivariate normal distribution limited application of this method until recently.

More recently extensions of the standard model have been proposed that maintain the hypothesis of fixed coefficients $\beta_i = \bar{\beta}$ but drop the assumption that the distribution of ε_i is known. Cosslett (1983), Manski (1985), Klein and Spady (1990) and Thompson (1989) proposed methods for estimating $\bar{\beta}$ in this "semiparametric" version of the model. The fixed coefficients $\bar{\beta}$ is also estimable using a variety of estimators for semiparametric regression models more general than the one considered here. See Han (1987), Ichimura (1991), Stoker (1986) or Ruud (1986) for examples.

This paper reunifies the random coefficients and semiparametric extensions to the standard binary choice model. By allowing for random coefficients we allow for more general forms of unobserved heterogeneity than

the standard model. On the other hand, the restrictions that we impose on the joint distribution of β_i and ε_i are the minimum needed to achieve identification of this distribution given the assumption that all unobservables are independent of x_i. So we avoid the potential for misspecification that arises when parametric functional forms are imposed on this distribution.

2. Model Identification

We assume that x_i includes a constant term: $x_i' = (1, \tilde{x}_i')$. This means that the first element of β_i is indistinguishable from ε_i in (1). We assume $\varepsilon_i = 0$ throughout in order to resolve this simple non-identifiability.

We assume that the cumulative distribution function of β_i is $F_0 \in \mathcal{F}$, where $\mathcal{F}$ is some space of distribution functions. Let $H(x) = \{ b: x'b \geq 0 \}$. Given that x_i and β_i are independent, model (1) requires that $\Pr(y_i=1 \mid x_i=x) = p_0(x) \equiv p(x,F_0)$, where

$$p(x,F) \equiv \int_{H(x)} dF. \tag{4}$$

The distribution F_0 is identified relative to $\mathcal{F}$ if and only if for each $F \in \mathcal{F}$,

$$\Pr\{ p(x_i,F) = p_0(x_i) \} = 1 \Rightarrow F = F_0 \tag{5}$$

If τ is any functional on $\mathcal{F}$ then we will say that $\tau_0 \equiv \tau(F_0)$ is identified (relative to $\mathcal{F}$) if and only if for each $F \in \mathcal{F}$

$$\Pr\{ p(x_i,F) = p_0(x_i) \} = 1 \Rightarrow \tau(F) = \tau_0 \tag{6}$$

These definitions are perfectly general since our independence assumption implies that all information about F_0 is contained in the distribution of y_i conditional on x_i.

It is well known that a scale normalization is needed for identification. In fact, $\mathcal{F}$ must be normalized against a very broad class of scale transformations of β_i:

Lemma 1. Let $g: \mathbb{R}^K \to (0,\infty)$ and let F be the distribution function of $g(\beta_i)\cdot\beta_i$. If $F \in \mathcal{F}$ then F_0 is not identified.

Proof: The conclusion follows because $x_i'\beta_i \geq 0 \Leftrightarrow g(\beta_i)\cdot x_i'\beta_i \geq 0$. ∎

Parametric binary choice models typically normalize the scale of β_i by fixing one diagonal element of $\mathrm{var}(\beta_i)$, then proceed to estimate some or all of the remaining first and second moments. Theorem 1 shows that this procedure is not sufficient to identify F_0 when $\mathcal{F}$ is sufficiently large.

Theorem 1. Let Σ be any fixed, symmetric, KxK positive semi-definite matrix. Suppose that $\Pr\{\ \beta_i = 0\ \} = 0$, that the distribution of $\beta_i/\|\beta_i\|$ is absolutely continuous with respect to (K-1)-dimensional Hausdorff measure on the unit sphere in $\mathbb{R}^K$,[2] and that the corresponding density function f is strictly positive. Then there exists a sequence of random vectors $\tilde{\beta}_n$ with distribution functions F_n such that

(i) $p(x,F_n) = p(x,F_0)$ for every x and n, and

(ii) $E(\tilde{\beta}_n) \to 0$ and $E(\tilde{\beta}_n\tilde{\beta}_n') \to \Sigma$ as $n \to \infty$.

Corollary. The correlation matrix of β_i is not identified.

Proof. Let $s_1,\ldots,s_K$ be an orthonormal set of eigenvectors for Σ and let $\lambda_1,\ldots,\lambda_K$ be the corresponding non-negative eigenvalues. Let $\{\varepsilon_n\}$ be a sequence of numbers in (0,.2) that converges to zero. Without loss of generality assume that $\|\beta_i\| = 1$. (Replace β_i with $\beta_i/\|\beta_i\|$ and apply Lemma 1 otherwise.) For each k and n let

$$C_{kn} = \{ \beta\colon \|\beta\| = 1,\ |s_k'\beta| > 1-\varepsilon_n \}$$

$$p_{kn} = \Pr\{ \beta_i \in C_{kn}\}$$

$$w_{kn} = \sqrt{\lambda_k/p_{kn}}$$

$$V_{kn} = E[\ \beta_i\beta_i' \mid \beta_i \in C_{kn}\ ]$$

First we develop some preliminary facts. The bound $\varepsilon_n < .2$ ensures that the sets C_{kn}, $k = 1,\ldots,K$ do not overlap. The assumptions made about the distribution of β_i and about ε_n ensure that $p_{kn} > 0$ and $w_{kn} \geq 0$ for all k and n, and that $p_{kn} \to 0$ for each k as $n \to \infty$. Thus $p_{kn}w_{kn} = \sqrt{\lambda_k p_{kn}} \to 0$ for each k as $n \to \infty$ as well. Adopt the matrix norm $\|A\| \equiv \|\mathrm{vec}(A)\|$ so that $\|\beta_i\beta_i'\| \leq \|\beta_i\|^2 = 1$. Then $\|V_{kn}\| \leq E[\ \|\beta_i\beta_i'\| \mid \beta_i \in C_{kn}\ ] \leq 1$. Likewise, setting $V \equiv E(\beta_i\beta_i')$ we must have $\|V\| \leq 1$. Since

$$\beta\beta' - s_k s_k' = (\beta - s_k)(\beta - s_k)' + (\beta - s_k)s_k' + s_k(\beta - s_k)'$$

and since $\|s_k\| = 1$, we have

$$\|\beta\beta' - s_k s_k'\| \leq \|\beta - s_k\|^2 + 2\ \|\beta - s_k\|.$$

Now if $\|\beta\| = 1$ and $s_k'\beta > 1-\varepsilon_n > .8$, an upper bound on $\|\beta - s_k\|$ is given by $\cos^{-1}(1-\varepsilon_n) < 1.3$ So the right-hand side of the preceding equation is bounded by $3\cdot\cos^{-1}(1-\varepsilon_n)$. Replace β with $-\beta$ in the derivation of this bound in order to deduce that the bound applies for all β in C_{kn}. Therefore

$$\|V_{kn} - s_k s_k'\| \le E[\ \|\beta_i \beta_i' - s_k s_k'\| \mid \beta_i \in C_{kn}\ ]$$

$$\le 3 \cos^{-1}(1-\varepsilon_n)$$

$$\to 0 \text{ as } n \to \infty.$$

We now construct the required sequence of random variables $\{\tilde{\beta}_n\}$. For each n let $\tilde{\beta}_n = g_n(\beta_i)\beta_i$, where

$$g_n(\beta) = \varepsilon_n + \sum_{k=1}^{K} w_{kn} 1\{\ \beta \in C_{kn}\ \}.$$

By construction we have $g_n > 0$, so conclusion (i) of the theorem holds by application of Lemma 1.

To verify the remaining conclusions, note first that $\|\tilde{\beta}_n\| = g_n(\beta_i)$, so

$$E\ \|\tilde{\beta}_n\| = E\ g_n(\beta_i) = \varepsilon_n + \sum_{k=1}^{K} w_{kn} p_{kn} \to 0 \text{ as } n \to \infty,$$

which is a sufficient condition for $E(\tilde{\beta}_n) \to 0$. Next, note that

$$\|E(\tilde{\beta}_n \tilde{\beta}_n') - \Sigma\| = \left\| E[g_n(\beta_i)^2 \beta_i \beta_i'] - \Sigma \right\|$$

$$= \left\| \varepsilon_n^2 V + 2\ \varepsilon_n \sum_{k=1}^{K} w_{kn} p_{kn} V_{kn} + \sum_{k=1}^{K} w_{kn}^2\ p_{kn} V_{kn} - \sum_{k=1}^{K} \lambda_k s_k s_k' \right\|$$

$$\le \varepsilon_n^2 + 2\ \varepsilon_n \sum_{k=1}^{K} w_{kn} p_{kn} + \sum_{k=1}^{K} \lambda_k \|V_{kn} - s_k s_k'\|$$

$$\to 0 \text{ as } n \to \infty.$$

The corollary follows immediately from the theorem since any covariance or correlation matrix satisfies the restrictions on Σ in the theorem. ■

The intuition behind Lemma 1 and Theorem 1 is that model (1) restricts the distribution of $\beta_i/\|\beta_i\|$ but does not restrict the distributions of $\|\beta_i\|$ conditional on $\beta_i/\|\beta_i\|$. We are free to rescale each possible value of β_i by a different positive constant. By "stretching" β_i in certain directions (corresponding to the eigenvectors of Σ in the proof of theorem 1) and "shrinking" β_i in other directions we can achieve an arbitrary covariance pattern without changing any of the conditional choice probabilities. The corollary to Theorem 1 shows that the correlation estimates reported in studies using standard parametric random coefficient models must be interpreted cautiously.

The preceding results suggest that a fairly stringent scale normalization is required if F_0 is to be identified. However, Theorem 2 shows that even when $\|\beta_i\| = 1$ is imposed it is often possible to construct many distributions for β_i that are observationally equivalent. So scale normalizations are not sufficient for identification.

Theorem 2. Let $B = \{ \beta : \|\beta\| = 1 \}$ and let μ be the (K-1)-dimensional Hausdorff measure on B. Suppose that $\mathcal{F}$ consists of all distributions that are absolutely continuous with respect to μ. Thus for each $F \in \mathcal{F}$, $Pr_F\{ \beta_i \in B \} = 1$ and $Pr_F\{ x'\beta = 0 \} = 0$ for every $x \neq 0$. Suppose further that f_0 is the density of F_0 with respect to μ and that $f_0 \geq \delta > 0$ everywhere on B. Then F_0 is not identified.

Proof. Let $h: B \to \mathbb{R}$ be any function satisfying (i) $|h| \le 1$, (ii) $h(\beta) = h(-\beta)$, and (iii) $\int_{B\cap H(x_0)} h(\beta)\, d\mu(\beta) = 0$ for some value of $x_0 \neq 0$. Then for any other $x \neq 0$, from (ii) and (iii) we have

$$\begin{aligned}
\int_{B\cap H(x)} h(\beta)\, d\mu(\beta) &= \int_B 1\{\, x'\beta > 0 \,\}\, h(\beta)\, d\mu(\beta) \\
&= \int_B 1\{\, x'\beta > 0,\ x_0'\beta > 0 \,\}\, h(\beta)\, d\mu(\beta) \\
&\qquad + \int_B 1\{\, x'\beta > 0,\ x_0'\beta < 0 \,\}\, h(-\beta)\, d\mu(\beta) \\
&= \int_B 1\{\, x'\beta > 0,\ x_0'\beta > 0 \,\}\, h(\beta)\, d\mu(\beta) \\
&\qquad + \int_B 1\{\, x'\beta < 0,\ x_0'\beta > 0 \,\}\, h(\beta)\, d\mu(\beta) \\
&= \int_B 1\{\, x_0'\beta > 0 \,\}\, h(\beta)\, d\mu(\beta) \\
&= 0.
\end{aligned}$$

So conditions (ii) and (iii) together imply that (iii) holds for any vector $x \neq 0$.

Now fix $\alpha > 1/\delta$ and let $F(t) = \int_{\{\beta \le t\}} [f_0(\beta) + h(\beta)/\alpha]\, d\mu(\beta)$. By construction $F \in \mathcal{F}$ and $p(x,F) = p_0(x)$ for every value of x, but $F = F_0$ only if $h = 0$. ∎

In order to obtain a specification for $\mathcal{F}$ that permits identification one can assume:

Model (1) holds with $x_i' = (1, \tilde{x}_i')$, $\beta_i' = (\theta_i', 1)$ and $\varepsilon_i = 0$. (A1)

The random vectors $\tilde{x}_i$ and θ_i are independent. (A2)

$\mathcal{F}$ consists of all distribution functions on $\mathbb{R}^{K-1} \times \{1\}$. (A3)

$$\Pr\{ \tilde{x}_i \in E \} > 0 \text{ for each open } E \subset \mathbb{R}^{K-1}. \tag{A4}$$

Theorem 3. If assumptions (A1), (A2), (A3) and (A4) hold then F_0 is identified relative to $\mathcal{F}$.

Proof. Suppose that $F \in \mathcal{F}$ and that

$$\Pr\{ p(x_i,F) = p_0(x_i) \} = 1 \tag{7}$$

By (A4) this holds only if $p(x,F) = p(x,F_0)$ for almost every x of the form $x = (\lambda',t)'$, where $\lambda \in \Lambda \equiv (\mathbb{R}-\{0\}) \times \mathbb{R}^{K-2}$ and $t \in \mathbb{R}$. For each x of this form and each $F \in \bar{\mathcal{F}}$ we have $p(x,F) = \Pr_F\{ -\lambda'\theta_i \le t \}$, which must be nondecreasing and right-continuous in t. So the distribution of $\lambda'\theta_i$ is the same under F and F_0 for almost every $\lambda \in \Lambda$. This implies that $E_F[h(\lambda'\theta_i)] = E_0[h(\lambda'\theta_i)]$ for every bounded and continuous function h, for almost every $\lambda \in \Lambda$, where E_F denotes expectation under F and E_0 denotes expectation under F_0.

Now let ϕ_F and ϕ_0 be the characteristic functions of $\lambda'\theta_i$ under F and F_0 respectively. To prove that $F = F_0$ it suffices to show that $\phi_F = \phi_0$ everywhere. Since $\phi_F(\lambda)$ and $\phi_0(\lambda)$ are expectations of bounded continuous functions we must have $\phi_F = \phi_0$ for almost every $\lambda \in \Lambda$. Since characteristic functions are continuous (a consequence of the Lebesgue dominated convergence theorem) this holds on all of Λ, and in fact on $\bar{\Lambda}$, the closure of Λ. But $\bar{\Lambda} = \mathbb{R}^{K-1}$. ■

Of the assumptions made in Theorem 3, only (A1) and (A4) merit further discussion. Condition (A1) normalizes the scale of β_i by requiring $\beta_{iK} = 1$.

This assumption requires *a priori* knowledge of the sign of one random coefficient, and is somewhat restrictive since it implies that $p_0(x_i)$ is monotone increasing in x_{iK}. A normalization of the form $|\beta_{iK}| = 1$ would be less restrictive, but would not be sufficient to prevent failures of identification of the form specified in Theorem 2.

Assumption (A4) is perhaps more troublesome. It requires that the distribution of $\tilde{x}_i$ has an absolutely continuous component with everywhere nonzero density. This rules out discrete or bounded random variables from $\tilde{x}_i$, and also prevents $\tilde{x}_i$ from including terms that are functionally related (such as interaction terms together with main effects). Perhaps these conditions can be weakened. Identification of F_0 under weaker conditions on x_i will most likely require stronger restrictions on $\mathcal{F}$, however.

3. Maximum Likelihood Estimation

We now consider the problem of constructing a maximum likelihood estimate of F_0 and describe the geometry of a particular class of estimates. We assume

$$(x_i, y_i),\ i = 1, 2, 3, \ldots \text{ are i.i.d.} \tag{A5}$$

Let $Z_i = (y_i, x_i)$. The conditional average log-likelihood function of a sample of N observations is

$$L(F) = \frac{1}{N}\sum_{i=1}^{N} \log f(Z_i, F) \tag{8}$$

where

$$f(Z_i, F) = p(x_i, F)^{y_i}\ [1-p(x_i, F)]^{(1-y_i)}. \tag{9}$$

(We set $\log f(Z_i,F) = -\infty$ whenever $f(Z_i,F) = 0$.) The maximum likelihood estimator $\hat{F}_N$ is defined to be any (measurable) solution to the equation

$$L(\hat{F}_N) = \max_{F\in\mathcal{F}} L(F). \tag{10}$$

Since $\mathcal{F}$ is generally an infinite-dimensional space, computation of $\hat{F}_N$ may appear intractable at first. It is perhaps not clear immediately that any solutions to (10) exist. In the remainder of this section we characterize the computational difficulties that must be overcome to compute an estimate $\hat{F}_N$ given the assumptions (A1) and (A3) of the preceding section. In particular we show that there always exists a solution $\hat{F}_N$ to (10) that is a discrete distribution with at most N points of support.

Let $B = \mathbb{R}^{K-1} \times \{1\}$, the support set in assumption (A3). The inequalities $x_i'b \geq 0$ and $x_i'b < 0$, $i \leq N$, partition B into a finite collection of sets $\mathcal{A}_N = \{ A_1, \ldots, A_M \}$. For each $F \in \mathcal{F}$, let q(F) be the M×1 vector of probabilities q(F) whose elements are given by

$$\lambda_j(F) = \int_{A_j} dF. \tag{11}$$

Since each $H(x_i)$ is a union of elements of $\mathcal{A}_N$, it follows that $p(x_i,F)$, hence $f(Z_i,F)$, is a sum of some of the elements of q(F). That is, for each observation there is an M×1 vector γ_i of zeros and ones for which $f(Z_i,F) = \gamma_i'q(F)$ for all $F \in \mathcal{F}$.

Clearly the range of q(F) is S_{M-1}, the unit simplex in $\mathbb{R}^M$. This permits problem (10) to be restated as

$$L(\hat{F}_N) = \max_{q\in S_{M-1}} \frac{1}{N}\sum_{i=1}^{N} \log \gamma_i'q. \tag{12}$$

The maximand on the right-hand side of (12) is concave and S_{M-1} is convex. There must exist at least one solution to the maximization problem,

which can be found using standard algorithms for concave programming subject to linear inequality constraints. Fletcher (1987) describes various algorithms for computing a solution.

The preceding discussion suggests a method for constructing a particular solution to (10). Let $\hat{q}$ be any solution to (12), and let t_j denote any point chosen arbitrarily from A_j, $j \le M$. Finally let

$$\hat{F}_N(t) = \sum_{j=1}^{M} \hat{q}_j \, 1\{ t_j \le t \}. \tag{13}$$

Substitution of (13) into (12) will verify that this $\hat{F}_N$ is a solution to the likelihood maximization problem (10) since we have $q(\hat{F}_N) = \hat{q}$. Obviously this solution is not unique.

Dimension Reduction Methods.

The dimension of each vector $q(F)$ is M. It can be shown[4] that $M = O(N^{K-1})$. Here we present two geometric arguments that enable one to restrict attention to problems of lower dimension. First, we show that any solution to (12) must have $\hat{q}_j = 0$ for a large subset of the indices $\{1,\ldots,M\}$ that can be determined *a priori*. So (12) can be replaced by an optimization over a lower-dimensional sub-simplex. Second, we show that there exists a solution to (12) in which at most N of the values $\hat{q}_j$ are non-zero. So a further reduction of (12) can be obtained by restricting attention to subsets of S_{M-1} that have this property.

To develop these ideas we introduce more notation. Let Γ be the N×M matrix of zeros and ones obtained by stacking the row vectors γ_i'. Then (12) can be rewritten as

$$L(\hat{F}_N) = \max_{g \in \mathcal{G}} \frac{1}{N} \sum_{i=1}^{N} \log g_i, \tag{14}$$

where $\mathcal{G}$ is the image of S_{M-1} under the linear transformation Γ.

Let λ_j denote the j'th column of Γ. We shall say that a point $b \in \mathbb{R}^K$ *correctly predicts* observation i if $y_i = 1\{ x_i'b \geq 0 \}$ and that A_j *correctly predicts* observation i if this holds for every $b \in A_j$. Similarly, λ_k *dominates* λ_j if $\lambda_j < \lambda_k$, which we take to mean $\lambda_j \leq \lambda_k$ (element by element) and $\lambda_j \neq \lambda_k$. Since the i'th element of λ_j is a binary indicator for whether or not A_j correctly predicts observation i, λ_k dominates λ_j if and only if the observations correctly predicted by A_j form a proper subset of those correctly predicted by A_k.

Theorem 4. Suppose that two columns of Γ satisfy $\lambda_j < \lambda_k$. Then $\hat{q}_j = 0$ for every solution $\hat{q}$ to (12).

Proof. The maximand in (14) is strictly monotone increasing on $\mathcal{G}$. Let $g = \Gamma q$ be a candidate solution to (14) for some $q \in S_{M-1}$ for which $q_j > 0$. Let $\tilde{q}$ coincide with q except that $\tilde{q}_j = 0$ and $\tilde{q}_k = q_j + q_k$. Let $\tilde{g} = \Gamma\tilde{q}$. Then $g < \tilde{g}$ since at least one element of λ_k exceeds the corresponding element of λ_j and all other elements coincide. But $\tilde{g} \in \mathcal{G}$ since $\tilde{q} \in S_{M-1}$. Conclude that the original q cannot solve (12). ∎

In order to make use of Theorem 4 one must identify the undominated columns λ_k so that the maximization in (12) can be restricted to the corresponding subsimplex of S_{M-1}. Identification of the undominated columns appears to require $O(M^2)$ vector comparisons of the form $\lambda_j < \lambda_k$. However,

this bound ignores the geometry of the problem. Suppose that $\lambda_j < \lambda_k$. Then there is a convex set of points $B_{jk} \subset \mathbb{R}^K$ that correctly predict all of the observations correctly predicted by A_j but none of the observations not correctly predicted by A_k. Obviously $A_j \subset B_{jk}$ and $A_k \subset B_{jk}$. So there is a path in B_{jk} leading from A_j to A_k. This path must intersect a region A_r adjacent to A_j. However $\lambda_j < \lambda_r$ if $A_r \subset B_{jk}$ by construction of the latter.

This shows that λ_j is dominated by a column corresponding to a region immediately adjacent to A_j, if it is dominated at all. Therefore undominated columns can be found by eliminating one of the columns in each pair of adjacent regions in $\mathcal{A}$. In the worst case this requires O(M) comparisons.

Our second result on dimensionality is given by Theorem 5, which is a particular case of a result due to Lindsay (1983) for a more general class of mixture models.[5]

Theorem 5. A solution $\hat{q}$ to (12) exists with at most N non-zero elements.

Proof. The set $\mathcal{G}$ is the convex hull of the columns of Γ. Therefore it is a convex polytope in $\mathbb{R}^N$. Furthermore, the maximand in (14) is strictly concave in g. So the solution vector $\hat{g}$ to (14) is unique, and lies on an exterior face of $\mathcal{G}$. Each exterior face of $\mathcal{G}$ is in turn the convex hull of a subset of the columns of Γ whose affine dimension is at most N-1. Then Carathéodory's Theorem (see Rockafellar, 1970, Theorem 17.1) requires existence of a vector $\hat{q} \in S_{M-1}$ with at most N non-zero elements satisfying $\hat{g} = \Gamma\hat{q}$. ■

4. Consistency of the ML Estimator

In order to prove that the maximum likelihood estimator of F_0 is consistent we will need some measure of distance between distribution functions in $\mathcal{F}$. The metric

$$d(F_1,F_2) = \| F_1 - F_2 \| \equiv \int |F_1(b) - F_2(b)| \, e^{-\|b\|} \, db \tag{15}$$

will do nicely.[6] Conditions are given below under which

$$\| \hat{F}_N - F_0 \| \xrightarrow{p} 0 \text{ as } N \to \infty \tag{16}$$

for each measurable version of $\hat{F}_N$.

Thompson (1989, Proposition A.1) shows that convergence in the metric (15) is equivalent to pointwise convergence at every continuity point of the limiting distribution function. So (15) induces a topology on $\mathcal{F}$ that is isomorphic to the topology of weak convergence of the corresponding probability measures. We conclude that whenever (16) is true and $\hat{\beta}_N$ is distributed according to $\hat{F}_N$ then $\hat{\beta}_N \xrightarrow{d} \beta_1$. In particular, (16) implies that $p(x,\hat{F}_N) \xrightarrow{p} p_0(x)$ for each fixed vector x satisfying $\Pr\{ x'\beta_1 = 0 \} = 0$, and that $E\, g(\hat{\beta}_N) \to E\, g(\beta_1)$ for every bounded, continuous function g.

It may happen that $\mathcal{F}$ is not compact under the given topology, which is inconvenient for the details of the forthcoming consistency proof. To avoid this possibility let $\bar{\mathbb{R}}^K$ denote the compact metric space obtained by embedding $\mathbb{R}^K$ into its unit ball via the mapping $\beta \to \beta/(1+\|\beta\|)$ and taking closure.[7] We assume

$\mathcal{F}$ consists of all distribution functions for probability measures with support contained in some compact set $C \subset \bar{\mathbb{R}}^K$. (A6)

Then compactness of $(\mathcal{F},d)$ follows from the isomorphism previously mentioned,[8] and from the fact that the set of all probability measures on a compact metric space is weakly compact (Parthasarathy, 1967, Theorem 6.4).

Assumption (A6) may appear restrictive at first since it is contradicted by (A3). The apparent restrictiveness is illusory. Suppose that we replace the set $\mathbb{R}^{K-1} \times \{1\}$ in (A3) with its closure with respect to $\bar{\mathbb{R}}^K$. Then the proof of Theorem 3 goes through, provided we also make the innocuous assumption $\Pr\{ \beta_i \in \mathbb{R}^K \} = 1$. The corresponding change needed in Section 3 is to replace each set $H(x_i)$ with its closure in $\bar{\mathbb{R}}^K$. This adds some points "at infinity" to some of the sets A_j, but is otherwise inconsequential: The geometric structure of $\hat{F}_N$ depends only on the indicator functions for these sets.

Before presenting the main result we establish some notation and further assumptions. Let $f(Z_i,F,\rho) = \sup_{\|F'-F\|<\rho} f(Z_i,F')$, with it understood that $F' \in \mathcal{F}$ always. Let $\mathcal{F}_0$ denote the subset of $\mathcal{F}$ that satisfies $\Pr_F\{ x'\beta_i = 0 \} = 0$ for every vector $x \neq 0$ and each $F \in \mathcal{F}_0$. Since $\mathcal{F}_0$ includes all continuous distributions on C, it is a dense subset of $\mathcal{F}$ with respect to the metric (15) whenever assumption (A6) holds. Assume:

$$\Pr\{ 0 < p_0(x_i) < 1 \} = 1. \tag{A7}$$

$$E \log f(Z_i,F_0) \text{ is finite.} \tag{A8}$$

$$\text{For each } F \in \mathcal{F}, \quad \| F - F_0 \| \neq 0 \;\Rightarrow\; E \log f(Z_i,F) < E \log f(Z_i,F_0). \tag{A9}$$

Since $f(Z_i,F) \leq 1$, $\log f(Z_i,F) \leq 0$ for every F. We can allow for the possibility $E \log f(Z_i,F) = -\infty$ in (A9) without any danger of confusion.

Identification of F_0 relative to $\mathcal{F}$ together with (A7) and (A8) implies (A9).

We will need the following Lemmas.

Lemma 2. For each $F \in \mathcal{F}_0$

$$\lim_{\rho \to 0} \log f(Z_i, F, \rho) = \log f(Z_i, F) \quad \text{almost surely.} \tag{17}$$

Proof. For given $F \in \mathcal{F}_0$, condition on the probability one event that

$$\int_{\{b:\ x_i' b = 0\}} dF = 0 \tag{18}$$

so that F is continuous on the boundary of $H(x_i)$. Suppose that $\lim_{\rho \to 0} \log f(Z_i, F, \rho) \neq \log f(Z_i, F)$. Then there exists a sequence $\{\rho_n\}$ such that $\rho_n \to 0$ and a sequence $\{F_n\} \subset \mathcal{F}$ such that $\| F_n - F \| < \rho_n$ and such that $\liminf_{n \to \infty} \log f(Z_i, F_n) > \log f(Z_i, F)$. But since F_n converges weakly to F we must have $p(x_i, F_n) \to p(x_i, F)$.[9] But $\log f(x_i, F)$ is a continuous function of $p(x_i, F_n)$, so we must have $\lim_{n \to \infty} \log f(Z_i, F_n) = \log f(Z_i, F)$, a contradiction. ∎

Lemma 3. For each $F \in \mathcal{F}_0$

$$\lim_{\rho \to 0} E \log f(Z_i, F, \rho) = E \log f(Z_i, F). \tag{19}$$

(Both sides may be $-\infty$.)

Proof. This proof follows the argument given in Theorem B, Section 27 of Halmos (1950, pp. 112-113). Suppose first that $E \log f(Z_i, F)$ is finite. Then since

$$0 \leq -\log f(Z_i, F, \rho) \leq -\log f(Z_i, F) \tag{20}$$

the Lebesgue dominated convergence theorem implies the result.

Next consider the case where $E \log f(Z_i, F) = -\infty$. We show that this implies $\lim_{\rho \to 0} E \log f(Z_i, F, \rho) = -\infty$. Suppose not. Then (since $E \log f(Z_i, F, \rho)$

is monotone decreasing as $\rho \downarrow 0$) $\lim_{\rho \to 0} E \log f(Z_i, F, \rho)$ exists and is finite. Let $\rho_n = 1/n$. Then for every $m, n > 0$, $\log f(Z_i, F, \rho_m) - \log f(Z_i, F, \rho_{m+n}) \geq 0$, which implies

$$E \mid \log f(Z_i, F, \rho_m) - \log f(Z_i, F, \rho_{m+n}) \mid$$
$$= E \log f(Z_i, F, \rho_m) - E \log f(Z_i, F, \rho_{m+n}) \to 0 \tag{21}$$

as $m, n \to \infty$. Thus $\{\log f(Z_i, F, \rho_n)\}$ is a Cauchy sequence with respect to the L_1 norm on $\mathbb{R}$. It must converge (in L_1) to an integrable function, and must have a subsequence that converges almost surely. By Lemma 2, the limit must be $\log f(Z_i, F)$. Thus $\log f(Z_i, F)$ is integrable, contradicting the hypothesis $E \log f(Z_i, F) = -\infty$. ∎

Finally we present the main result of this section. Theorem 6 establishes (weak) consistency for the maximum likelihood estimator $\hat{F}_N$ using an approach due to Wald (1949) as modified by Wolfowitz (1949).

Theorem 6. Fix $\delta > 0$ and let $\mathcal{F}(\delta) = \{ F \in \mathcal{F}: \| F - F_0 \| \geq \delta \}$. Given assumptions (A6), (A5), (A7), (A8) and (A9), there exists $\eta(\delta)$ with $0 < \eta(\delta) < 1$ such that

$$\Pr \left\{ \frac{\sup_{F \in \mathcal{F}(\delta)} \prod_{i=1}^{N} f(Z_i, F)}{\prod_{i=1}^{N} f(Z_i, F_0)} > \eta(\delta)^N \right\} \longrightarrow 0 \text{ as } N \to \infty. \tag{22}$$

Corollary. Under these same conditions, $\| \hat{F}_N - F_0 \| \xrightarrow{P} 0$ as $N \to \infty$.

Proof. For each $F \in \mathcal{F}_0 \cap \mathcal{F}(\delta)$ there exists $\rho_F > 0$ such that $E \log f(Z_i, F, \rho_F) < E \log f(Z_i, F_0)$. (This follows from Lemma 3 and assumption (A9).) Since $\mathcal{F}_0$ is dense in $\mathcal{F}$ there is a covering of $\mathcal{F}(\delta)$ consisting of open balls centered at points $F \in \mathcal{F}_0$ and radii ρ_F. Since $\mathcal{F}(\delta)$ is compact there exists a finite subcover. Let $(F_1, \rho_1), \ldots, (F_M, \rho_M)$ index a finite subcover. Note that

$$\prod_{i=1}^{N} f(Z_i, F) \le \sum_{m=1}^{M} \prod_{i=1}^{N} f(Z_i, F_m, \rho_m). \tag{23}$$

Thus

$$\Pr\left\{ \frac{\sup_{F \in \mathcal{F}(\delta)} \prod_{i=1}^{N} f(Z_i, F)}{\prod_{i=1}^{N} f(Z_i, F_0)} > \eta(\delta)^N \right\} \le \Pr\left\{ \sum_{m=1}^{M} \frac{\prod_{i=1}^{N} f(Z_i, F_m, \rho_m)}{\prod_{i=1}^{N} f(Z_i, F_0)} > \eta(\delta)^N \right\}$$

$$\le \sum_{m=1}^{M} \Pr\left\{ \frac{\prod_{i=1}^{N} f(Z_i, F_m, \rho_m)}{\prod_{i=1}^{N} f(Z_i, F_0)} > \frac{\eta(\delta)^N}{M} \right\}$$

$$\le \sum_{m=1}^{M} \Pr\left\{ \frac{1}{N} \sum_{i=1}^{N} \left[\log f(Z_i, F_m, \rho_m) - \log f(Z_i, F_0) \right] > \log \eta(\delta) - \frac{\log M}{N} \right\} \tag{24}$$

Define $\psi_m(Z_i) = \log f(Z_i, F_m, \rho_m) - \log f(Z_i, F_0)$. Then the last expression in (24) equals

$$\sum_{m=1}^{M} \Pr \left\{ \frac{1}{N} \sum_{i=1}^{N} \left[\psi_m(Z_i) - E\, \psi_m(Z_i) \right] > -E\, \psi_m(Z_i) + \log \eta(\delta) - \frac{\log M}{N} \right\} . \tag{25}$$

Since $-E\, \psi_m(Z_i) > 0$ for $m \le M$, for $\eta(\delta)$ close enough to one, for all $m \le M$ and all N large enough it is true that $-E\, \psi_m(Z_i) + \log \eta(\delta) - \log M / N > \varepsilon$ for some $\varepsilon > 0$. So the conclusion of the theorem follows from (24) by applying a law of large numbers to each of the sums inside the braces in (25). ■

As a final remark on consistency, we note that the i.i.d. sampling assumption (A5) is not necessary. It could be replaced by any condition that permits application of a weak law of large numbers in the last step of the proof of Theorem 6.

5. Computational Examples and Monte Carlo Evidence

We have already discussed some computational aspects of our estimator. Here we elaborate further in the context of a specific parameterization, and present some example calculations. We also summarize results from some Monte Carlo experiments investigating the performance of our estimator and of a parametric rival in some small sample problems.

Model Specifications.

Here and in the remainder of the paper we restrict attention to models satisfying assumptions (A1) - (A4) with K = 3. Thus $\mathcal{F}$ is the set of all distributions on $\mathbb{R}^2 \times \{1\}$. Although $\mathcal{F}$ is formally a set of distributions on a subset of $\mathbb{R}^3$, for simplicity we equate each distribution $F \in \mathcal{F}$ with its

marginal distribution with respect to the first two random coefficients and ignore the fixed third coefficient. For convenience we now let β_i denote the i-th component of the random coefficient vector.

We also restrict attention to cases with N = 1,000. The corresponding value of M depends on the random configuration of the x_i vectors. The vectors $\tilde{x}_i$ in our examples have the independent bivariate standard normal distribution. An upper bound is $M \leq 500,501$. This bound is achieved (with probability one) whenever $\tilde{x}_i$ has a continuous distribution, as in our examples.

Here and in the next section we consider three different specifications for F_0. These are summarized in Table 1. Model 1 takes F_0 as the bivariate standard normal distribution function. We consider this model an important test case, since the probit maximum likelihood estimator of the mean and covariance matrix for this model figures prominently in the applied literature. Models 2 and 3, represent two different kinds of departures from normality. Their density functions are depicted in Figures 1 and 2. The distribution of model 2 is characterized by a long, curved ridge, with two modes separated by a slightly lower saddle point. The iso-density contours are shaped roughly like a boomerang, leading to a nonlinear dependency among β_1 and β_2. Model 3, like model 1, preserves independence of β_1 and β_2, but has a distinct bimodality. We feel that detection of features like those present in models 2 and 3 provide a strong motivation for our new estimator. The parameter values selected for these models were chosen subjectively in order to achieve the approximate shapes appearing in the figures.

Computational Algorithms.

To test the feasibility of our estimator we have written a computer program that calculates the nonparametric maximum likelihood estimator of F_0.[10] Our program solves the geometry problem described in section 3, and eliminates the parameters corresponding to dominated columns of Γ as described in that section. We then choose an initial starting value for q in problem (12) using a procedure described below. Problem (12) is solved using the BFGS concave maximization algorithm with exact line searches. The inequality constraints defining S_{M-1} are handled using an active sets method. These techniques are discussed extensively in Fletcher (1987). Since the constraint set is convex and the objective function concave, the algorithm converges monotonically to a solution with at most finitely many changes in the active set of constraints, and with no cycling possible.

The choice of starting values is very important due to the high dimensionality of the problem.[11] We have found the following algorithm to be quite efficient at finding a starting value close to the final solution. We start by sorting the undominated columns of Γ into decreasing order by their numbers of non-zero entries. We then select columns from this list, in order, if they satisfy the following criteria: Each newly selected column must have a one in at least one row for which all of the already selected columns have zeros. Selection terminates when a set is obtained such that there is no row i for which every member of the included set has a zero in position i. This ensures that any strictly convex combination of the included columns will not have any zero elements, which in turn ensures that the initial value of the log-likelihood is finite. The actual starting value is obtained by choosing a

q vector that weights each of the selected vectors proportionally to it's number of non-zero elements, with zero weight on all other columns of Γ.

This procedure is guaranteed to produce a starting value for q with at most N non-zero elements. This allows considerable simplification of the coding for the subsequent maximization problem, since it ensures that there will never be more than N unconstrained parameters during the successive searches. Keeping this dimension low is critical, since fast searches require approximation of an inverse Hessian matrix with dimension equal to the number of currently unconstrained parameters.

For comparison purposes, we also calculate the parametric maximum likelihood estimates corresponding to model 1 with unknown parameters μ and Σ. This probit ML estimator is calculated by choosing starting values of $\mu = 0$ and $\Sigma = I$ (the true values for model 1), and applying a commercial maximization subroutine (using analytical first and second derivatives) to the random coefficients probit likelihood. The estimated value of Σ was constrained slightly away from its region of singularity to avoid numerical instability of these calculations.

The estimation codes were written largely in Cray Pascal and were compiled and run on the Cray-2 supercomputer made available to us by the Minnesota Supercomputer Institute. Our code was able to make extensive use of the large memory and high precision vector processing capabilities of this machine.

Example Calculations.

The flavor of the estimation procedures can be obtained from examination of figures 3, 4 and 5. These figures give estimates obtained from three data

sets, each with 1,000 observations simulated from one of the models of table 1. Panels (a) and (b) of each figure present the nonparametric ML estimates as a pseudo-density function obtained by convoluting the estimated (discrete) distribution with a small amount of continuously distributed noise. Panels (c) and (d) of each figure display the corresponding density estimates produced by the probit estimator. Although the asymptotic theory presented in the last section does not support any formal interpretation for the nonparametric density estimates, we feel that these views of the estimates capture their important features quite well.

Examining figure 3, one finds (as expected) that the parametric ML estimate is closer to the true distribution than is the nonparametric estimate in the sample obtained from model 1. After all, the parametric ML estimate is efficient even among parametric estimators for this model. While the nonparametric estimate has too much fine variation, it does pick up the gross features of the bivariate standard normal distribution. The distribution estimated by our nonparametric procedure is roughly unimodal and fairly symmetric.

The probit estimate has relatively less of an advantage when applied to the data set used to produce figure 4. The probit likelihood does not allow any adaptation to the boomerang shaped iso-density contours of model 2. In contrast, the nonparametric estimate does pick up this feature to some extent. Again, however, the nonparametric estimate displays a great deal of roughness at finer scales.

The nonparametric estimate clearly outperforms the probit estimate in figure 5. It picks up the bimodality of model 3, and low variation around each mode, very well. The probit estimator does a reasonable job of locating

the center of the distribution of β_1 here. However, it badly underestimates the variance of β_1 and overestimates the variance of β_2 in a futile attempt to adapt to the bimodality of the true distribution.[12]

Monte Carlo Evidence.

The remainder of this section reports results from a Monte Carlo experiment designed to partially evaluate our proposed estimator. For each of the three models of table 1 we generated 500 simulated data sets of 1,000 observations each and applied both our nonparametric maximum likelihood estimator and the probit random coefficients maximum likelihood estimator to each. The results of these experiments are presented in the remaining tables and figures.

Quantiles of the Monte Carlo sampling distributions of $p(x,\hat{F}_N)$ are summarized in tables 2—4. The tables also present the true values of $p_0(x)$, and the bias[13] and root-mean-squared error (RMSE) for estimating $p_0(x)$. There are two rows in each table for each value of x. These present results for the nonparametric and probit maximum likelihood estimators respectively. Figures 6—8 display surface plots of the RMSE figures for each estimator and each model on a finer grid.

Table 2 and figure 6 show that both estimators estimate the choice probabilities in model 1 quite well. RMSE for the nonparametric estimator never exceeds .07, while the bias for this estimator never exceeds .03. The probit estimator, as expected, displays lower RMSE numbers (≤.05) and considerably lower biases (≤.01). There is a noticeable dip in the RMSE for both estimators in the vicinity of the points (0,-2) and (0,2). These are points where the true probabilities are very close to zero and one

respectively. RMSE for the probit estimator increases noticeably in the corners of the grid in figure 6, where the density of observations is relatively low. The prediction performance of the two estimators is closest at these points.

Turning next to table 3 and figure 7 we see that the probit estimator does not dominate the nonparametric estimator in model 2. The worst case RMSE for the probit estimator is .085 at the origin. The bias of the probit estimator at the origin is .081, so bias is the main component of prediction error at this point. The worst case RMSE for the nonparametric estimator is only slightly worse here than in model 1, reaching a peak of .078 at the point (2,0). In contrast to the probit estimator, the bias of .010 is a minor component of the nonparametric estimator prediction error at this point. The worst case bias of the nonparametric estimator is .028 at the point (-2,2). As in model 1, both estimators have good predictive performance in the vicinity of the points (0,-2) and (0,2) where the true probabilities are very close to zero and one.

Evidence in table 4 suggests that the nonparametric estimator has better predictive performance for most x values in model 3. The worst case bias of the nonparametric estimator in this model is .05 while the worst case bias of the probit estimator is .170. The bias of the probit estimator consistently exceeds .04 except along the line $x_2 = 0$ (where the probit estimator is asymptotically unbiased) and at the point (1,-2) where the true value is effectively zero. In contrast, the bias of the nonparametric estimator exceeds .04 at only six of 25 points listed in table 4. The RMSE figures for the probit estimator are dominated by their bias components, reaching a peak of .171 at the points (-2,0) and (2,0) where the probit bias is at a maximum.

The RMSE for the nonparametric estimator peaks at these same points, but at a much lower value of .073. The RMSE of the nonparametric estimator appears to be less dominated by bias than is the RMSE of the probit estimator here.

Figure 8 clearly shows large regions (corresponding to true probabilities close to zero and one) where the predictive performance of the nonparametric estimator is very good. On the other hand, the RMSE of the probit estimator exhibits large swings. The ridges in the right-hand panel of the figure correspond to regions of high bias for the probit estimator.

As a measure of the relative prediction efficiency of the two estimators we calculated the quantity RPE = $\log_{10}(RMSE_{NP}/RMSE_P)$ on a grid of (x_2,x_3) values, where $RMSE_{NP}$ and $RMSE_P$ denote the root-mean-squared error for predicting $p_0(x)$ for the nonparametric and probit estimators respectively. These numbers, truncated to the range [-1,1], are displayed in figures 9—11. (The truncation is necessary since the computed ratio is numerically unstable at x values for which $p_0(x)$ is very close to zero or one due to the limited number of Monte Carlo trials.) The positive and negative values of RPE are plotted separately in figures 10 and 11 as an aid to interpretation. Positive values of RPE correspond to greater prediction efficiency for the probit estimates, while negative values correspond to a greater prediction efficiency for the nonparametric estimator.

Figure 9 reflects the superior performance of the probit maximum likelihood estimator when its assumed model is true. The probit estimator has its greatest advantage along the line $x_3 = 0$, where $p_0(x) = 0.5$. Figures 10 and 11 show that neither estimator uniformly dominates the other when the probit model is false. Figure 10 reveals that that the probit estimator has a better RPE for most x values in model 2. The probit estimator performs

dramatically better than the nonparametric estimator near the points (.5,2) and (1,-2) but dramatically worse at the point (.5,2). However, these are all points at which the density of the x values is low and at which the absolute performance of both estimators is extremely good. Overall the probit estimator probably has a small advantage for most purposes, although we would expect this advantage to decrease in larger samples.

In contrast, figure 11 suggests that the nonparametric estimator has much better predictive performance than the probit estimator in model 3 for very many x values. The exceptions occur along three ridges clearly evident in the upper panel where the probit model is coincidentally asymptotically unbiased.[14] There is also some evidence of superior probit performance along a very narrow ridge along the line $x_2 = 0$, where the probit estimator is also asymptotically unbiased. However, the probit estimator is inconsistent for $p_0(x)$ over a large portion of the grid. In contrast to the situation in model 2, not all of these points correspond to small absolute prediction errors for the probit model, and some of them occur at points where the density of x is relatively large.

Table 5 displays some summary statistics on the computational requirements for the nonparametric estimator encountered in our Monte Carlo simulations. Clearly this estimator requires substantial resources to calculate, and would have been infeasible on most economics research budgets until recently. However, the calculations are feasible on the current generation of mainframe computers and advanced workstations.

The remainder of the table demonstrates the effectiveness and importance of our strategies for dimension reduction. Recall that we have M = 500,501 in each of these simulations. Elimination of dominated regions lowered the

dimension of the calculations by roughly 95 percent. The maximum number of active (i.e. non-zero) parameters is typically encountered at the starting value for each estimation. Our algorithm for choosing starting values kept this number below 200 in every simulation. The final solutions never had more than 54 active parameters, which is roughly .01 percent of M, and considerably lower than the worst-case upper bound of N = 1,000.

6. Conclusion

This paper presents a maximum likelihood estimator of a binary choice model with random coefficients. The model does not require parametric specification of the distribution of the coefficients. We discussed identification and proved consistency under suitable conditions. We showed that the estimator does a reasonable job of recovering the unknown coefficient distribution in some small example problems. In our Monte Carlo experiments the new estimator performed well. The nonparametric estimator has superior predictive performance relative to the probit estimator when the true coefficient distribution is significantly non-normal, due to the bias displayed by the probit estimator. The new estimator requires substantial computer resources to calculate, but is feasible using current technology.

We have left a number of important questions for future research. First, the asymptotic theory for our estimator is incomplete. We present no rates of convergence for functionals of the estimated distribution, nor do we

develop a formal theory for testing parametric against nonparametric specifications for the random coefficients model.

Second, we feel that stronger identification results can be achieved when there is stronger a priori information about the coefficient distribution than we make use of here. For example, it seems likely that our assumption (A4) could be weakened if it were known a priori that certain of the coefficients were fixed rather than random. Pursuit of this extension would yield a synthesis of our estimator, which allows all coefficients (except one) to be random, with the estimator proposed by Cosslett (1983), which is a special case of our estimator requiring all coefficients (except for the intercept) to be fixed.

Finally, the model and method described here are generalizable to settings involving polytomous choice. Extensions to panel data are also possible in which full independence between the regressors and coefficients can be partially relaxed.

ACKNOWLEDGEMENT

We have benefited from communications with Charles Geyer, John Geweke and Herman Rubin. We thank the Minnesota Supercomputer Institute for providing computer support. Professor Thompson thanks the National Science Foundation for support provided by grant SES—9110419.

REFERENCES

Albright, Richard L., Steven R. Lerman and Charles F. Manski (1977). *Report on the Development of an Estimation Program for the Multinomial Probit Model.* Report for the Federal Highway Administration. Cambridge Systematics, Inc.: Cambridge, Massachusetts.

Billingsley, Patrick (1968). *Convergence of Probability Measures.* John Wiley & Sons: New York.

Billingsley, Patrick (1986). *Probability and Measure, Second Edition.* John Wiley and Sons: New York.

Cosslett, Stephen R. (1983). "Distribution-free maximum likelihood estimator of the binary choice model." *Econometrica* 51, 765-782.

Fletcher, R. (1987). *Practical Methods of Optimization, Second Edition.* John Wiley & Sons: Chichester.

Halmos, Paul R. (1950). *Measure Theory.* D. Van Nostrand: New York.

Han, Aaron K. (1987). "A non-parametric analysis of transformations." *Journal of Econometrics* 35, 191-209.

Hausman, Jerry A., and David A. Wise (1978). "A conditional probit model for qualitative choice: discrete decisions recognizing interdependence and heterogeneous preferences." *Econometrica* 46, 403-426.

Ichimura, Hidehiko (1991). "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models." Discussion Paper No. 264. Center for Economics Research, Department of Economics, University of Minnesota, Minneapolis.

Klein, Roger W., and Richard H. Spady (1990). "An efficient semiparametric estimator for discrete choice models." Bellcore Economics Discussion Paper #67.

Lindsay, Bruce G. (1983). "The geometry of mixture likelihoods: a general theory." *The Annals of Statistics* 11, 86-94.

Manski, Charles F. (1985). "Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator." *Journal of Econometrics* 27, 313-333.

McFadden, Daniel (1989). "A method of simulated moments for estimation of discrete response models without numerical integration." *Econometrica* 57, 995-1026.

Parthasarathy, K. R. (1967). *Probability Measures on Metric Spaces.* Academic Press: New York.

Pollard, David (1984). *Convergence of Stochastic Processes*. Springer-Verlag: New York.

Rockafellar, R. Tyrrell (1970). *Convex Analysis*. Princeton University Press: Princeton, New Jersey.

Ruud, Paul A. (1986). "Consistent estimation of limited dependent variable models despite misspecification of distribution." *Journal of Econometrics* 32, 157-187.

Stoker, Thomas M. (1986). "Consistent estimation of scaled coefficients." *Econometrica* 54, 1461-1481.

Thompson, T. Scott (1989). "Least Squares Estimation of Semiparametric Discrete Choice Models." mimeograph. University of Minnesota Department of Economics, Minneapolis.

Wald, Abraham (1949). "Note on the consistency of the maximum likelihood estimate." *The Annals of Mathematical Statistics* 20, 595-601.

Wolfowitz, J. (1949). "On Wald's proof of the consistency of the maximum likelihood estimate." *The Annals of Mathematical Statistics* 20, 601-602.

FOOTNOTES

1 These papers allow for more than two alternatives in the choice set.

2 See Billingsley (1986, Section 19) for an explanation of this measure. It is a natural generalization of (K-1)-dimensional Lebesgue measure to the surface of a sphere.

3 Take $\cos^{-1}$ in the interval $[0,\pi]$ here.

4 This bound can be obtained through application of Theorem II.16 and Lemma II.18 in Pollard (1984). A precise bound is given by

$$M \le \sum_{k=1}^{K} \binom{N}{k-1}.$$

This bound is attained if every set of x_i vectors taken K at a time are linearly independent.

5 The problem of finding $\hat{q}$ to solve (12) corresponds to finding a maximum likelihood estimate of a mixing distribution under the hypothesis that F_0 is a mixture of point distributions. This is because λ_j gives the vector of likelihoods for the individual observations associated with a distribution F that is degenerate at the point t_j.

6 Obviously $d(F_1,F_2)$ is symmetric in its arguments and satisfies the triangle inequality. Furthermore, $d(F_1,F_2) = 0$ implies that $F_1 = F_2$ almost everywhere. But since distribution functions are monotone and continuous from above, this implies that $F_1 = F_2$. So $d(\cdot,\cdot)$ is a metric on $\mathcal{F}$.

7 This is equivalent to adding a point "at infinity" in each direction of recession from the origin of $\mathbb{R}^K$. The topology can be metrized using any norm on the unit ball applied to the embedded points.

8 A function F is a distribution function for a probability measure on $\bar{\mathbb{R}}^K$ if and only if there is a probability measure μ such that $F(t) = \mu(\ cl\{\beta: \beta \le t\}\ )$, where "cl" denotes closure with respect to $\bar{\mathbb{R}}^K$. Technically d is not a metric on $\mathcal{F}$ given this extension since we can have $d(F_1,F_2) = 0$ for two distribution functions corresponding to probability measures that agree on $\mathbb{R}^K$ but not on $\bar{\mathbb{R}}^K$. So formally, $\mathcal{F}$ should be considered as a space of equivalence classes of distribution functions. The distinction is irrelevant for the present discussion since we will only be considering neighborhoods of F_0, and $d(F,F_0) = 0$ requires $F = F_0$, since F_0 is a proper distribution function on $\mathbb{R}^K$.

9 This follows from the Portmanteau Theorem. See Billingsley (1968, page 11).

10 The current version of our program explicitly handles datasets that might have been generated by models 1-3, that is, with K = 3 and one scale normalization, yielding an effective two-dimensional coefficient distribution. The code exploits some features of this setting that are not available in higher dimensional problems. A more general code can be written, but at the sacrifice of some computational speed when applied to the two-dimensional models considered here.

11 The first test version of our program simply started at the vector q whose elements were all equal to 1/M. We found that the minimization algorithm then spent enormous amounts of time successively activating constraints until it found a much lower dimensional problem close to the solution. The algorithm for choosing starting values described here produces several orders of magnitude improvement in the performance of our algorithm relative to this first attempt.

12 In fact, the variance of β_1 in the distribution estimated by the probit model in figure 5 was even lower than that displayed. The actual distribution that was estimated corresponds to a degenerate ridge rather than a proper density. We have added a little bit of noise to the panel in figure 5 in order to produce a displayable plot.

13 Although the tables present the signs and magnitudes of the biases, our discussion will use the term bias to refer to the magnitudes only.

14 The three ridges correspond to the three points at which the cumulative distribution function for a univariate normal distribution intersects the cumulative distribution function for a sharply bimodal mixture of normal distributions with the same overall scale.

Table 1. Simulated Distributions for $\beta = (\beta_1, \beta_2)'$

Model 1:	$\beta \sim \mathcal{N}(\ \mu,\ \Sigma\ ),\ \mu = 0,\ \Sigma = I$
Model 2:	β is an equal weighted mixture of γ_1 and γ_2

$$\gamma_1 \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ -\mu \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right)$$

$$\gamma_2 \sim \mathcal{N}\left(\begin{bmatrix} -\mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_2^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix} \right)$$

$$\mu = 0.3587,\ \sigma_1^2 = 0.26271,\ \sigma_2^2 = 0.06568,\ \text{and}\ \rho = -0.1$$

Model 3:	β_1 and β_2 independently distributed

$$\beta_1 \sim \mathcal{N}(\ 0,\ \sigma^2\ )$$

β_2 an equally weighted mixture of normals γ_1 and γ_2

$$\gamma_1 \sim \mathcal{N}(\ 0.2806,\ \sigma^2\ ),\ \gamma_2 \sim \mathcal{N}(\ -1.6806,\ \sigma^2\ )$$

$$\sigma^2 = 0.038462$$

Table 2. Monte Carlo sampling distributions for $p(x,\hat{F}_N)$ in Model 1.

				Quantiles				
(x_2,x_3)	$p_0(x)$	Bias	RMSE	.10	.25	.50	.75	.90
(-2,-2)	.18555	-.01491	.06378	.09144	.12787	.17147	.21375	.25057
		-.00376	.05042	.11861	.14928	.18326	.21503	.24588
(-2,-1)	.32736	-.01674	.06930	.22637	.26431	.30798	.35663	.40177
		-.00485	.04309	.26851	.29680	.32783	.35063	.37238
(-2,0)	.50000	-.00379	.06123	.41514	.45556	.49647	.53736	.57036
		-.00012	.02756	.46381	.48105	.50111	.51767	.53501
(-2,1)	.67264	.01066	.06647	.59970	.63897	.68102	.72830	.77149
		.00478	.04004	.63190	.65042	.67447	.70141	.72796
(-2,2)	.81445	.01555	.06565	.75188	.78650	.82783	.87132	.92256
		.00406	.04748	.75952	.78469	.81504	.84853	.88080
(-1,-2)	.07865	-.00750	.05213	.00000	.02951	.06747	.10281	.14359
		.00174	.02944	.04373	.05909	.07795	.09778	.11858
(-1,-1)	.23975	-.01619	.06153	.14216	.18010	.22264	.26647	.29809
		-.00104	.03591	.18826	.21427	.23979	.26155	.28282
(-1,0)	.50000	.00073	.05674	.42491	.46116	.49997	.54034	.57777
		.00013	.02531	.46642	.48348	.50157	.51797	.53284
(-1,1)	.76025	.01334	.06083	.69711	.73266	.77398	.81356	.84971
		.00137	.03385	.71799	.73897	.75988	.78464	.80499
(-1,2)	.92135	.01248	.04992	.87179	.90108	.93486	.97498	.99999
		-.00135	.02756	.88211	.90181	.92270	.94085	.95316
(0,-2)	.02275	-.01047	.02513	.00000	.00000	.00000	.01617	.04873
		.00106	.01157	.01003	.01484	.02241	.03079	.03907
(0,-1)	.15866	-.02867	.06187	.05969	.09371	.12876	.16809	.19904
		-.00134	.02777	.12109	.13758	.15850	.17798	.19155
(0,0)	.50000	-.00089	.06090	.42183	.46096	.49798	.53354	.57649
		.00035	.02167	.47169	.48615	.50046	.51438	.52709
(0,1)	.84134	.02395	.06056	.79380	.82552	.86352	.90605	.93686
		.00159	.02887	.80562	.82318	.84388	.86265	.87751

Table 2 (continued).

(x_2,x_3)	$p_0(x)$	Bias	RMSE	Quantiles .10	.25	.50	.75	.90
(0,2)	.97725	.00892	.02742	.95073	.98111	1.00000	1.00000	1.00000
		-.00108	.01180	.96052	.96893	.97792	.98509	.98943
(1,-2)	.07865	-.01150	.04916	.00000	.03124	.06311	.10144	.13328
		.00105	.02908	.04493	.05832	.07716	.09721	.11813
(1,-1)	.23975	-.01614	.05905	.15116	.18182	.22533	.26155	.29722
		-.00183	.03496	.19265	.21494	.23838	.26120	.28087
(1,0)	.50000	.00079	.05815	.42577	.46212	.50083	.54086	.57298
		-.00011	.02478	.46842	.48449	.50027	.51675	.52899
(1,1)	.76025	.01450	.06174	.69735	.73740	.77765	.81473	.85052
		.00169	.03446	.71807	.73777	.76224	.78615	.80480
(1,2)	.92135	.01061	.05091	.86531	.90118	.93546	.97314	1.00000
		-.00106	.02863	.88316	.90128	.92307	.94067	.95541
(2,-2)	.18555	-.01555	.06378	.09219	.12952	.16864	.21721	.24988
		-.00466	.04926	.11769	.14978	.18503	.21418	.24014
(2,-1)	.32736	-.01059	.06717	.22801	.27258	.31837	.36100	.40452
		-.00544	.04177	.26905	.30053	.32460	.35121	.36933
(2,0)	.50000	-.00186	.06552	.41252	.45483	.49770	.54295	.58286
		-.00002	.02772	.46542	.48234	.49986	.51862	.53285
(2,1)	.67264	.01234	.06627	.59877	.63656	.69085	.73335	.76643
		.00533	.04274	.62933	.65053	.67181	.70506	.73618
(2,2)	.81445	.01368	.06315	.74773	.78446	.83062	.87244	.90633
		.00448	.05017	.75726	.78342	.81516	.85154	.88409

Note: The first and second row of each pair report results for the nonparametric and probit maximum likelihood estimators respectively.

Table 3. Monte Carlo sampling distributions for $p(x,\hat{F}_N)$ in Model 2.

				Quantiles				
(x_2,x_3)	$p_0(x)$	Bias	RMSE	.10	.25	.50	.75	.90
(-2,-2)	.05721	.00631	.04691	.00000	.02821	.06021	.09502	.13202
		.03689	.04795	.05915	.07499	.09304	.11184	.13096
(-2,-1)	.28370	-.02374	.07083	.17319	.21193	.25811	.30664	.34774
		-.02865	.04645	.21111	.23281	.25511	.27783	.29932
(-2,0)	.54067	-.02574	.06778	.43414	.47333	.51592	.56244	.59138
		-.03493	.04436	.47076	.48502	.50625	.52556	.54104
(-2,1)	.73435	.01328	.06413	.66287	.70938	.74826	.79209	.82234
		.01977	.04007	.71243	.73060	.75386	.77641	.79690
(-2,2)	.90175	.02831	.05658	.86122	.89761	.93160	.97037	1.00000
		.00896	.02975	.87725	.89295	.91064	.93074	.94442
(-1,-2)	.00782	.00167	.01838	.00000	.00000	.00000	.01047	.03628
		.01457	.02015	.00968	.01402	.02036	.02710	.03554
(-1,-1)	.15981	-.01841	.05561	.07392	.10289	.14150	.17665	.21156
		-.00475	.02358	.11970	.13448	.15421	.17294	.19302
(-1,0)	.50000	-.00420	.05946	.42064	.45452	.49303	.53283	.57346
		.00434	.02592	.47027	.48608	.50503	.52270	.53660
(-1,1)	.84019	.02524	.06321	.79082	.82901	.86711	.90356	.93909
		.01014	.02885	.81742	.83385	.85076	.86867	.88385
(-1,2)	.99218	.00471	.01136	.99216	1.00000	1.00000	1.00000	1.00000
		-.01317	.01768	.96667	.97405	.98059	.98676	.99064
(0,-2)	.00034	-.00025	.00123	.00000	.00000	.00000	.00000	.00000
		-.00006	.00104	.00002	.00005	.00012	.00026	.00603
(0,-1)	.05273	-.02360	.03994	.00000	.00000	.02332	.05017	.07116
		-.01764	.02340	.01814	.02484	.03320	.04210	.05479
(0,0)	.41940	.02694	.07043	.36720	.40442	.44235	.48932	.52473
		.08073	.08480	.46509	.48219	.50253	.51801	.53162
(0,1)	.99491	.00295	.00878	1.00000	1.00000	1.00000	1.00000	1.00000
		-.02959	.03255	.94886	.95796	.96733	.97454	.98024

Table 3 (continued).

(x_2,x_3)	$p_0(x)$	Bias	RMSE	Quantiles .10	.25	.50	.75	.90
(0,2)	1.00000	.00000	.00000	1.00000	1.00000	1.00000	1.00000	1.00000
		-.00025	.00069	.99948	.99972	.99988	.99995	.99998
(1,-2)	.00014	.00093	.00570	.00000	.00000	.00000	.00000	.00000
		.00016	.00057	.00001	.00002	.00011	.00037	.00076
(1,-1)	.03443	-.00700	.03264	.00000	.00000	.01890	.04584	.07254
		-.00057	.01790	.01356	.02001	.03200	.04519	.05745
(1,0)	.50000	.01892	.07642	.38273	.43391	.48449	.53048	.57820
		-.00771	.03871	.44191	.46667	.49502	.51621	.53944
(1,1)	.96557	.01225	.03291	.92790	.96084	1.00000	1.00000	1.00000
		-.00196	.01797	.93919	.95261	.96515	.97684	.98552
(1,2)	.99986	-.00031	.00313	1.00000	1.00000	1.00000	1.00000	1.00000
		.00058	.00059	.99913	.99962	.99988	.99997	.99999
(2,-2)	.02805	-.00952	.02867	.00000	.00000	.00000	.03172	.06353
		.02547	.01726	.00372	.00944	.01819	.03093	.04299
(2,-1)	.14558	-.01206	.05937	.06665	.09018	.12651	.17313	.20906
		-.00137	.04610	.08008	.11158	.14438	.17822	.20088
(2,0)	.46847	.00972	.07827	.37563	.42195	.47850	.52996	.58018
		.02197	.04349	.44286	.46532	.49257	.51277	.53668
(2,1)	.86529	-.00090	.06236	.78763	.81932	.86595	.90844	.94258
		-.01905	.04807	.79110	.81736	.84607	.87269	.90571
(2,2)	.99020	-.00563	.02649	.94330	.97389	1.00000	1.00000	1.00000
		-.01377	.02174	.95492	.96662	.98005	.98907	.99452

Note: The first and second row of each pair report results for the nonparametric and probit maximum likelihood estimators respectively.

Table 4. Monte Carlo sampling distributions for $p(x,\hat{F}_N)$ in Model 3.

				Quantiles				
(x_2,x_3)	$p_0(x)$	Bias	RMSE	.10	.25	.50	.75	.90
(-2,-2)	.49952	-.05039	.06915	.38902	.41814	.45257	.48365	.50550
		-.12955	.13303	.33129	.35002	.37043	.39168	.40866
(-2,-1)	.50009	-.00052	.04098	.44554	.47393	.49940	.52804	.55093
		.04944	.04971	.51910	.53313	.55050	.56628	.58001
(-2,0)	.55017	.04431	.07220	.52520	.55398	.59090	.63066	.67174
		.16931	.17059	.69221	.70543	.71881	.73427	.74690
(-2,1)	.92076	.01930	.05290	.87154	.90626	.94378	.98547	1.00000
		-.07080	.07359	.82337	.83584	.85070	.86430	.87653
(-2,2)	.99974	.00012	.00181	1.00000	1.00000	1.00000	1.00000	1.00000
		-.06823	.07003	.90992	.92146	.93277	.94233	.95087
(-1,-2)	.06236	-.01529	.06091	.00000	.00000	.01583	.08313	.13096
		.04786	.05539	.07565	.09207	.10871	.12814	.14601
(-1,-1)	.49647	-.04752	.06939	.38172	.41740	.45112	.48442	.51196
		-.12571	.12955	.33122	.35044	.37154	.39184	.40956
(-1,0)	.57793	.03067	.06797	.53484	.56215	.60723	.64526	.69122
		.13968	.14156	.68852	.70122	.71803	.73375	.74731
(-1,1)	.99763	.00146	.00558	1.00000	1.00000	1.00000	1.00000	1.00000
		-.06793	.06996	.90738	.91852	.93102	.94165	.95035
(-1,2)	1.00000	.00000	.00000	1.00000	1.00000	1.00000	1.00000	1.00000
		-.00947	.01077	.98374	.98809	.99137	.99422	.99616
(0,-2)	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000
		.00000	.00000	.00000	.00000	.00000	.00000	.00000
(0,-1)	.00000	-.00000	.00000	.00000	.00000	.00000	.00000	.00000
		.00000	.00003	.00000	.00000	.00000	.00000	.00001
(0,0)	.50000	-.00226	.11161	.35050	.42572	.49206	.56899	.63579
		-.01320	.14582	.36342	.44259	.49867	.55032	.60372
(0,1)	1.00000	.00000	.00000	1.00000	1.00000	1.00000	1.00000	1.00000
		-.00000	.00002	.99999	1.00000	1.00000	1.00000	1.00000

Table 4 (continued).

(x_2,x_3)	$p_0(x)$	Bias	RMSE	Quantiles .10	.25	.50	.75	.90
(0,2)	1.00000	.00000	.00000	1.00000	1.00000	1.00000	1.00000	1.00000
		.00000	.00000	1.00000	1.00000	1.00000	1.00000	1.00000
(1,-2)	.00000	-.00000	.00000	.00000	.00000	.00000	.00000	.00000
		.00949	.01085	.00415	.00584	.00858	.01167	.01572
(1,-1)	.00237	-.00178	.00470	.00000	.00000	.00000	.00000	.00000
		.06786	.06982	.04994	.05891	.06931	.07985	.09068
(1,0)	.42207	-.02866	.06684	.31173	.35156	.39352	.43624	.47323
		-.14030	.14207	.25155	.26686	.28193	.29789	.30968
(1,1)	.50353	.04656	.06807	.48769	.51498	.54577	.58489	.61533
		.12440	.12846	.58493	.60689	.62879	.64876	.67061
(1,2)	.93764	.01544	.06386	.87091	.91726	.98592	1.00000	1.00000
		-.04894	.05697	.84887	.87276	.89130	.90911	.92248
(2,-2)	.00026	-.00014	.00196	.00000	.00000	.00000	.00000	.00000
		.06819	.06996	.04998	.05754	.06775	.07736	.08898
(2,-1)	.07924	-.01918	.05333	.00000	.01590	.05541	.09356	.12784
		.07066	.07337	.12527	.13614	.15065	.16327	.17512
(2,0)	.44983	-.04389	.07251	.32585	.36732	.41085	.44898	.47883
		-.16962	.17086	.25267	.26577	.28078	.29529	.30624
(2,1)	.49991	-.00045	.03979	.44825	.47282	.49922	.52533	.55125
		-.04996	.05518	.41917	.43303	.45013	.46595	.48036
(2,2)	.50048	.04961	.06852	.49117	.51538	.54662	.58195	.61515
		.12890	.13248	.58925	.60932	.62963	.64950	.66961

Note: The first and second row of each pair report results for the nonparametric and probit maximum likelihood estimators respectively.

Table 5. Summary of computational requirements for the nonparametric estimator.

	Model 1		Model 2		Model 3	
	Median	Maximum	Median	Maximum	Median	Maximum
Undominated regions (out of 500,501)	24,230	27,816	13,668	17,674	13,073	16,172
Maximum active parameters during iterations	138	174	102	128	74	109
Active parameters at the solution	37	54	32	43	24	34
Processor time required (seconds)	108.7	160.3	56.8	84.6	45.8	63.5

Note: These statistics are from 500 Monte Carlo trials on each model.

Figure 1. True coefficient density for Model 2.

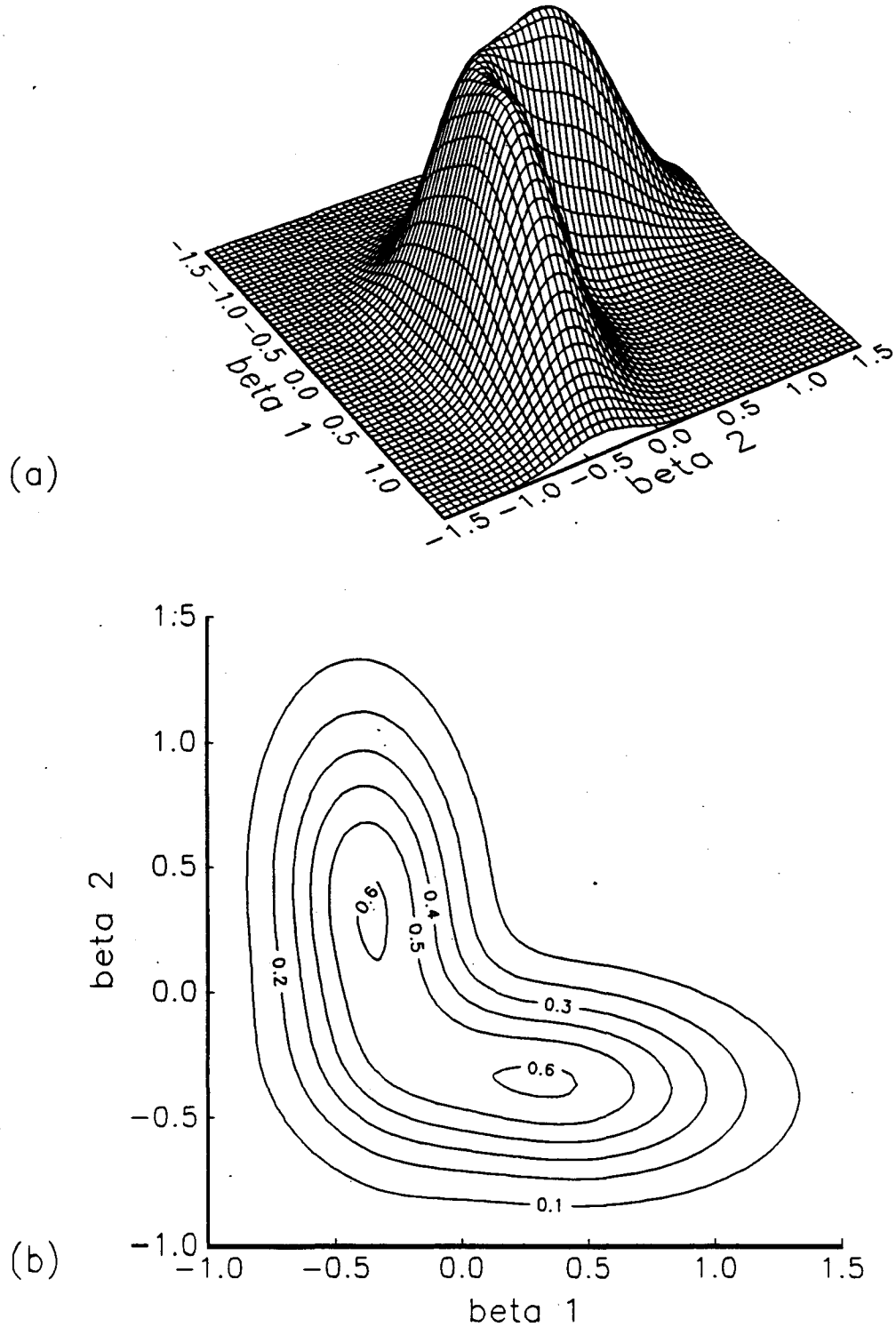


Figure 2. True coefficient density for Model 3.

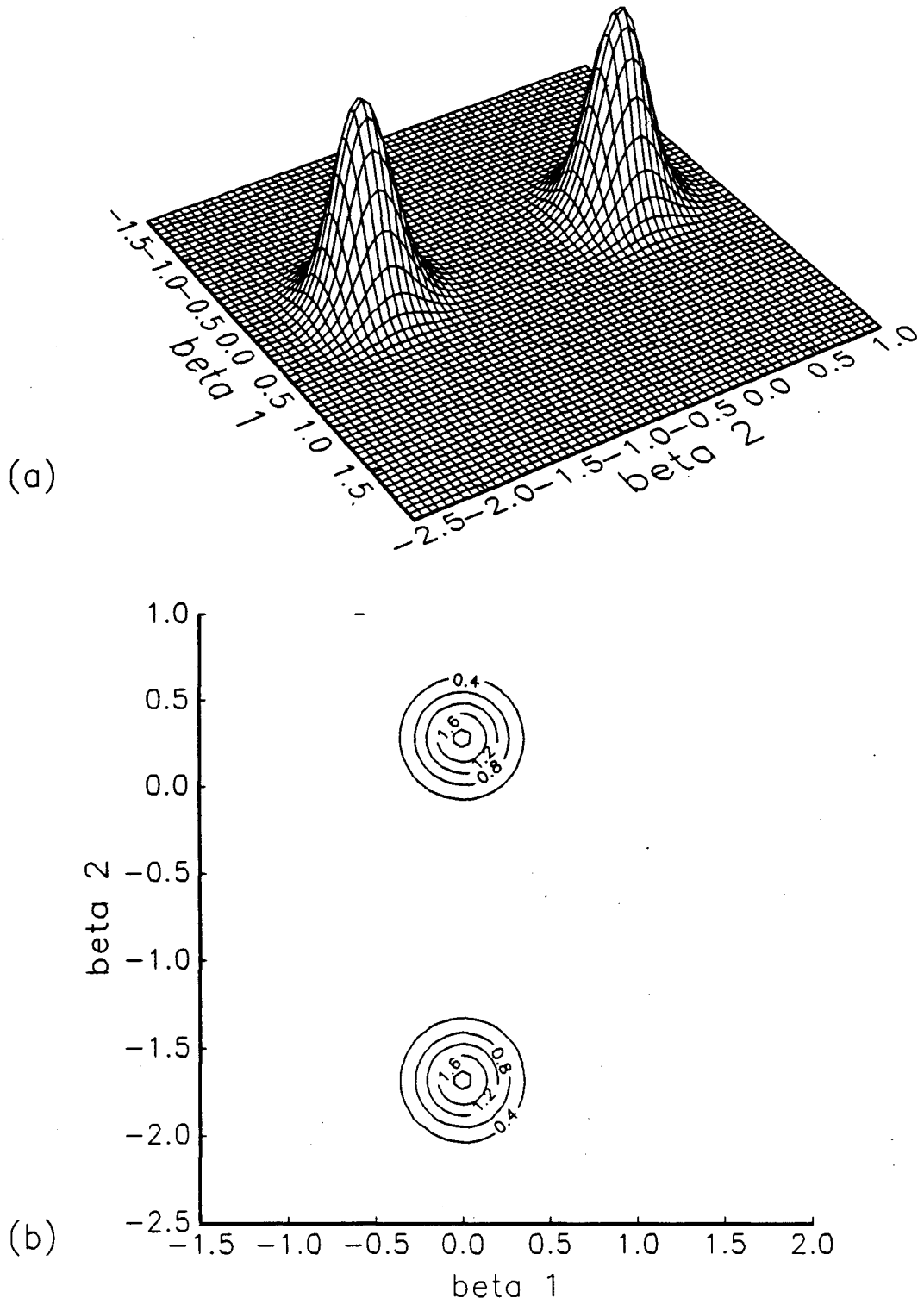
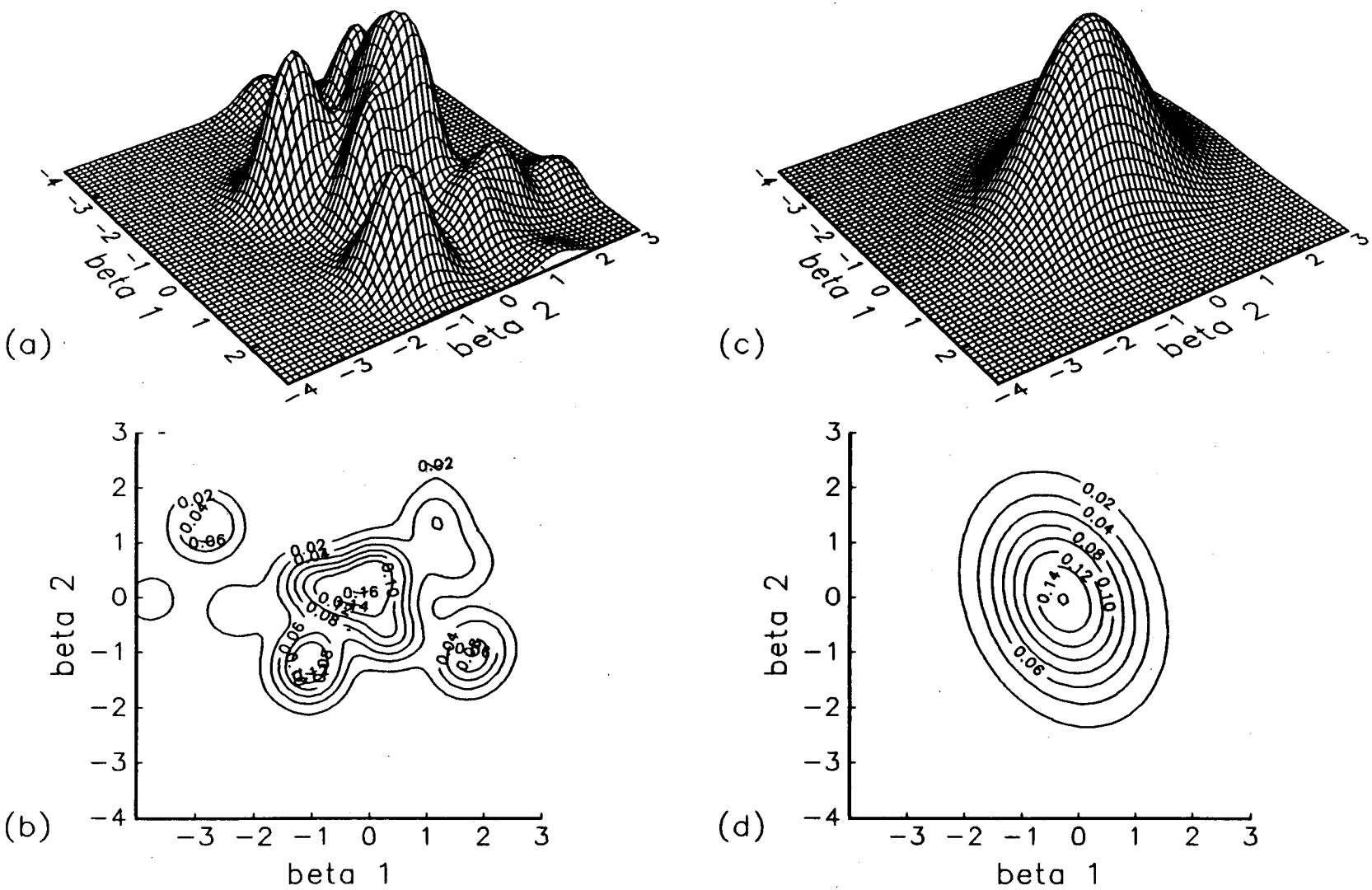
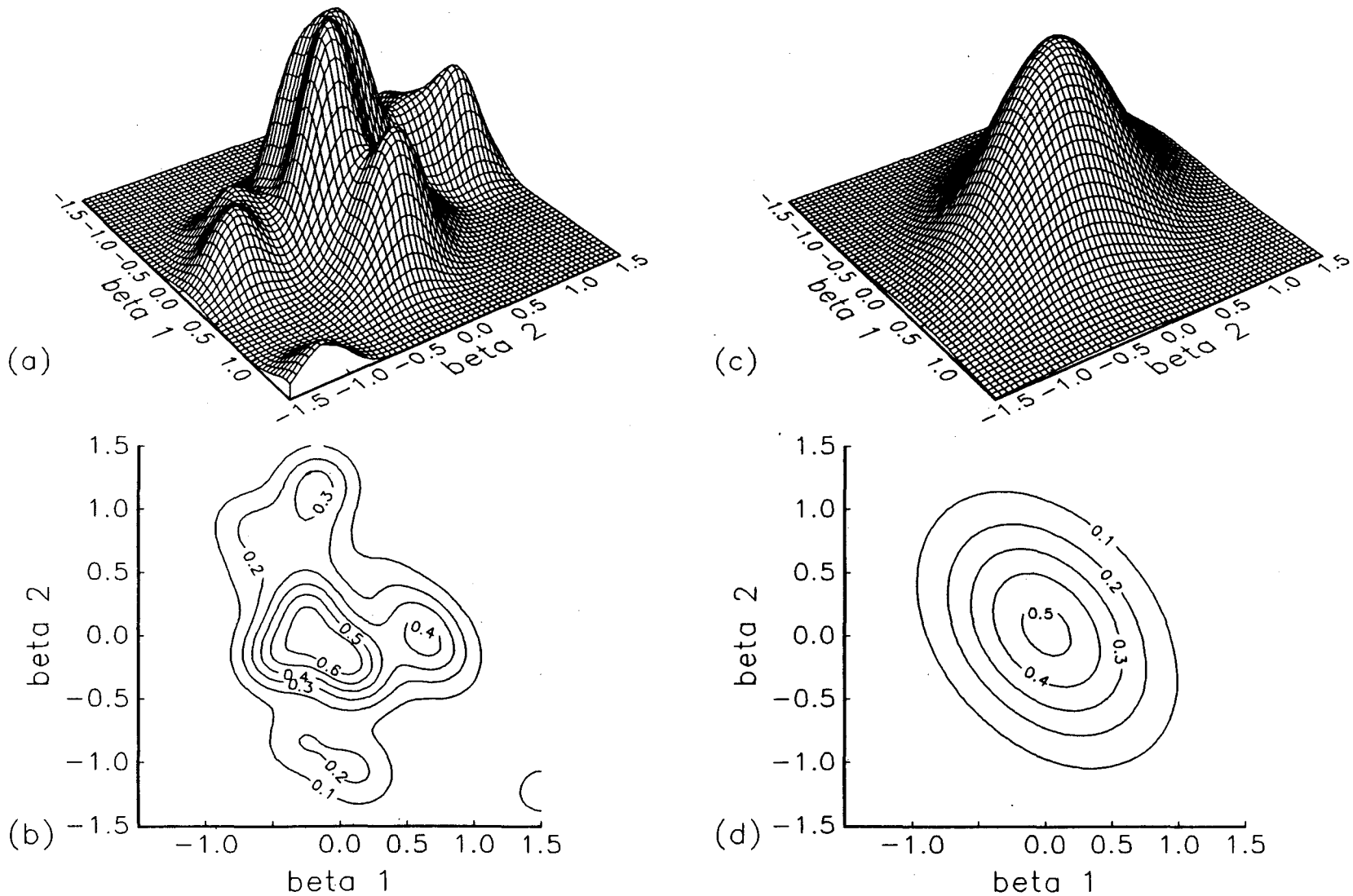


Figure 3. Example estimates for model 1.



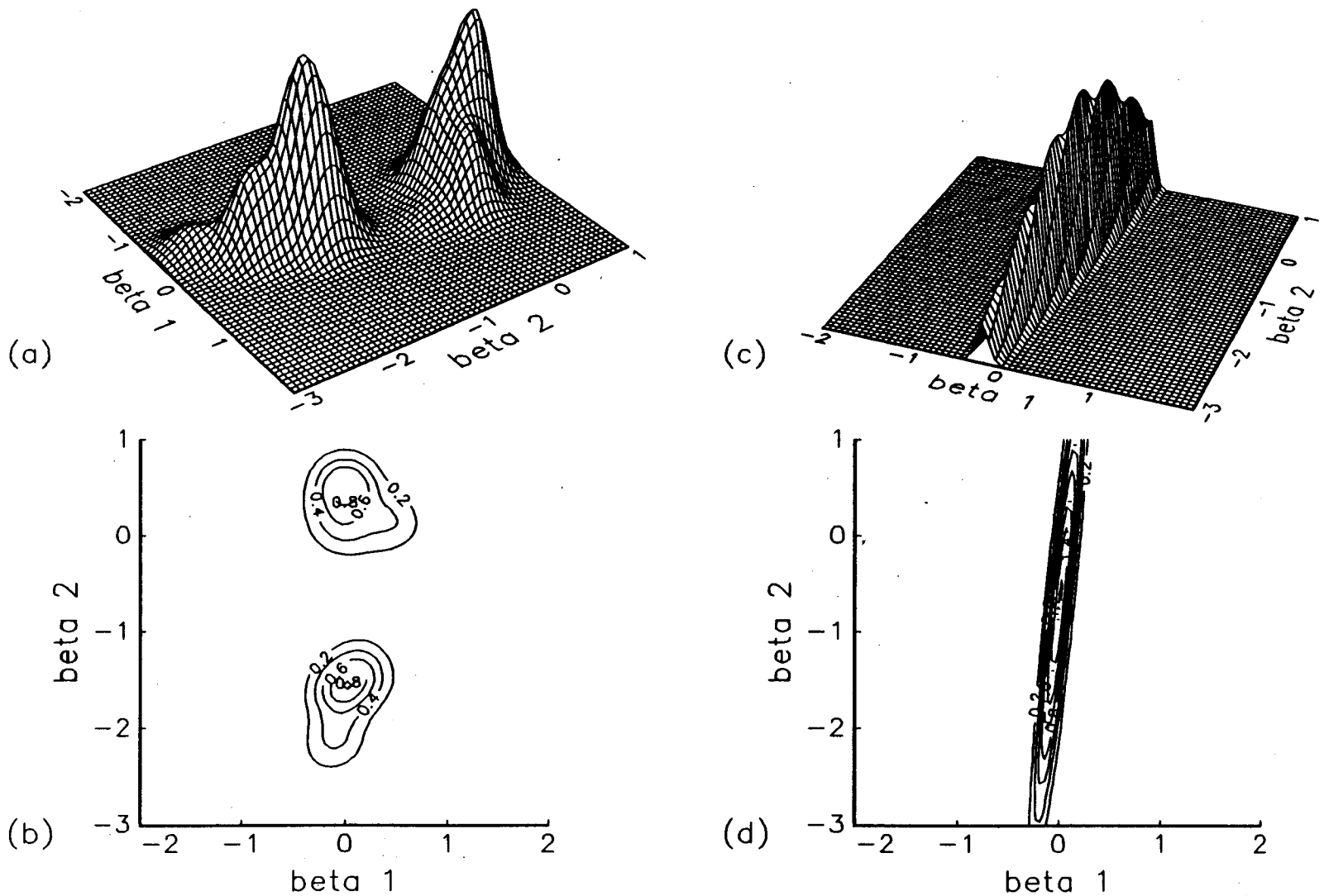
Note: Panels (a) and (b) are nonparametric estimates; (c) and (d) are probit estimates.

Figure 4. Example estimates for model 2.



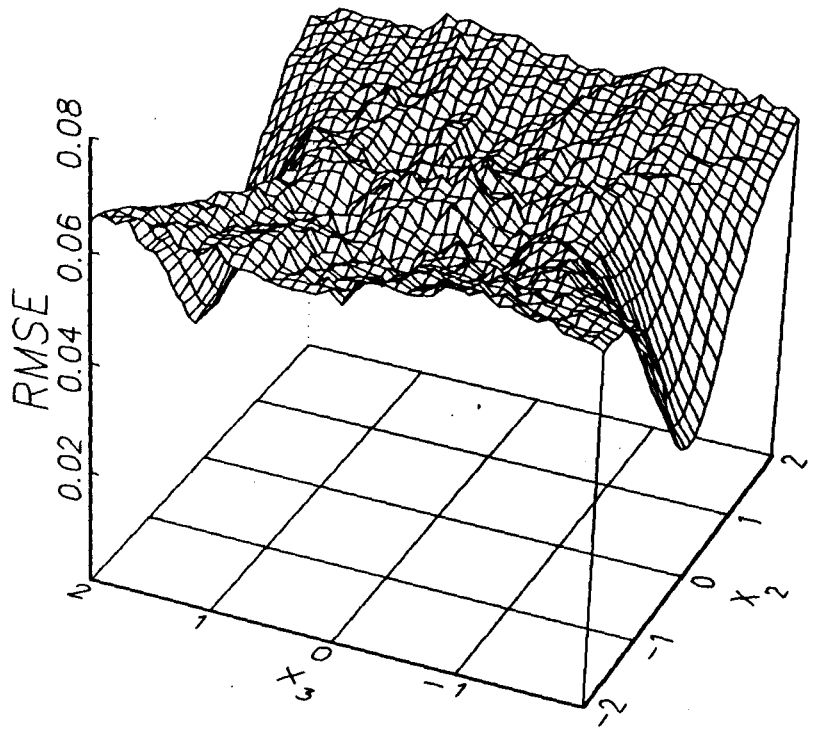
Note: Panels (a) and (b) are nonparametric estimates; (c) and (d) are probit estimates.

Figure 5. Example estimates for model 3.

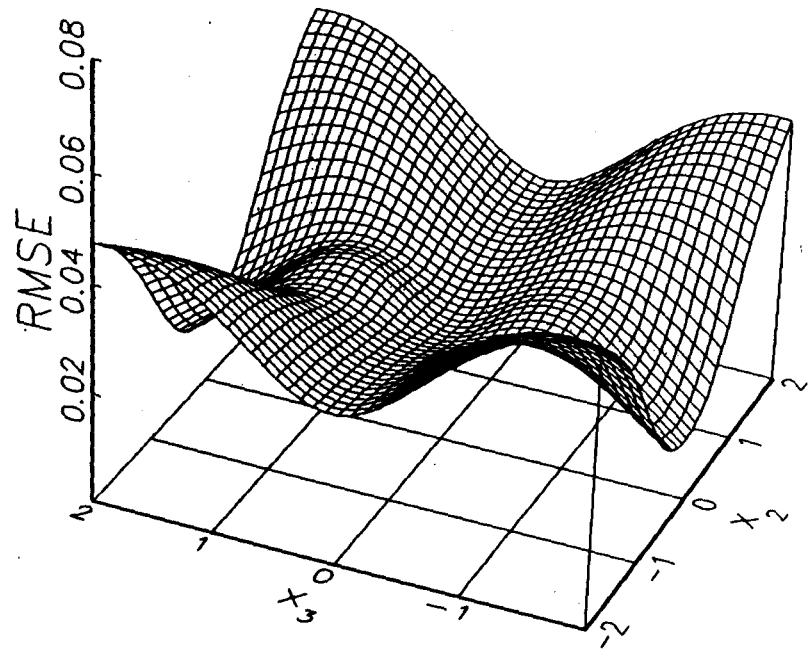


Note: Panels (a) and (b) are nonparametric estimates; (c) and (d) are probit estimates.

Figure 6. Estimated Root Mean Square Prediction Errors in Model 1.

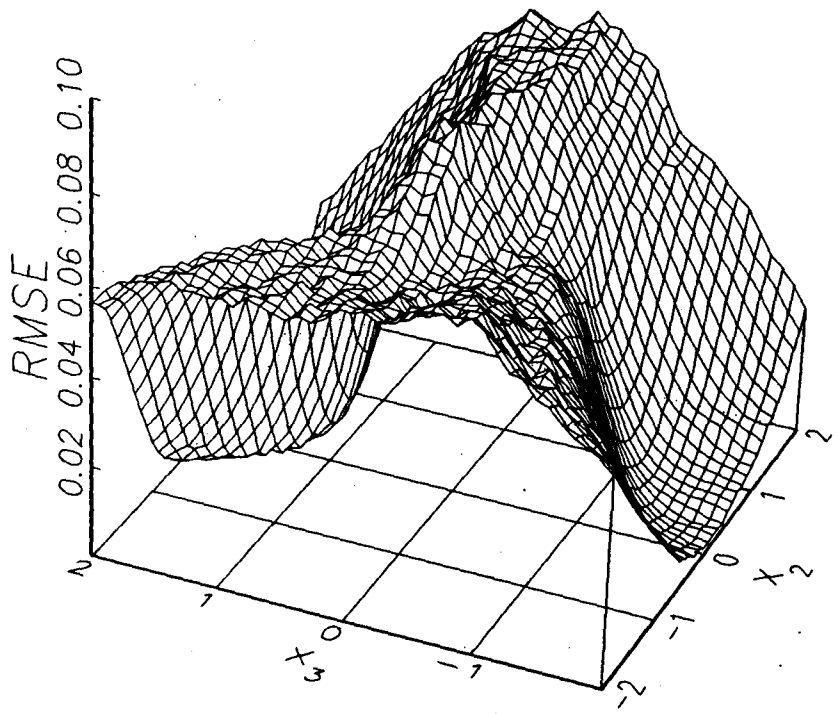


Nonparametric

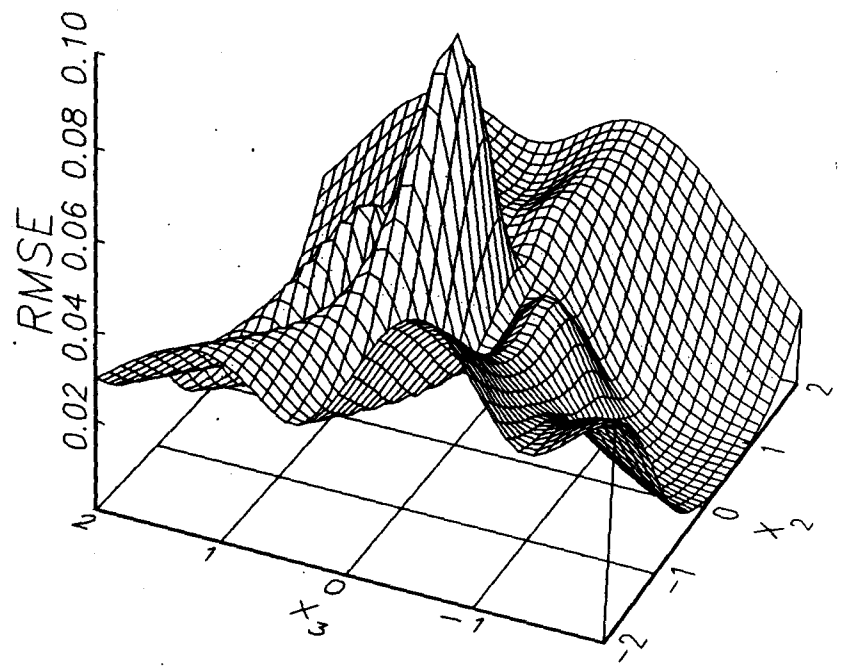


Probit

Figure 7. Estimated Root Mean Square Prediction Errors in Model 2.

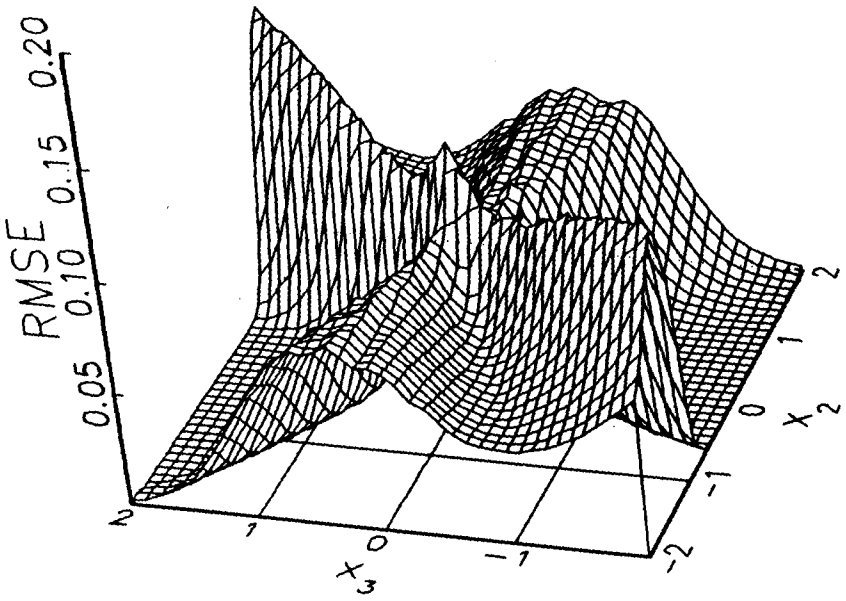


Nonparametric

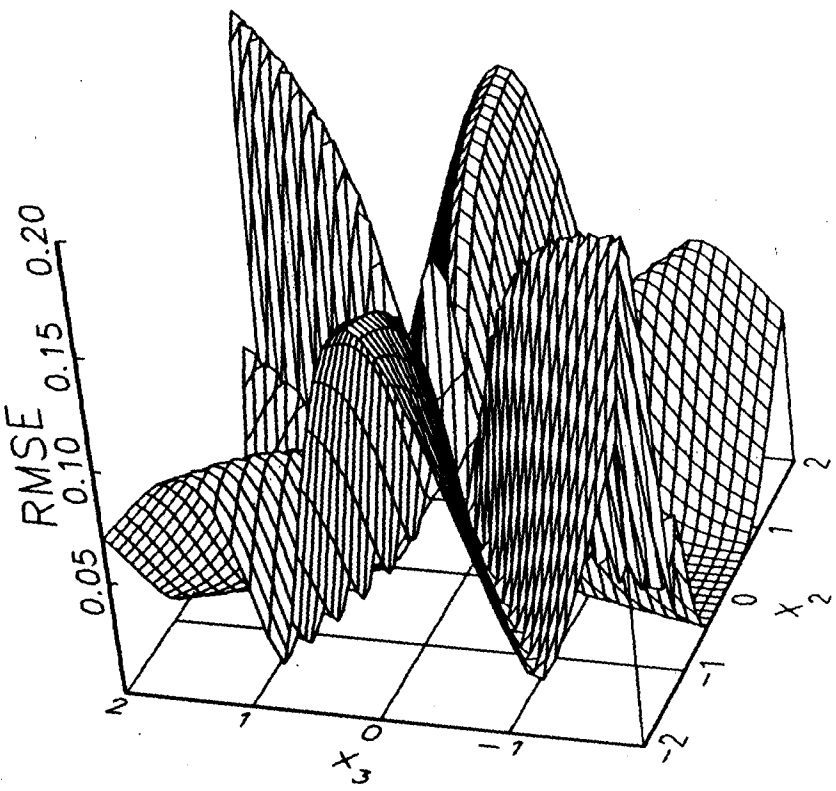


Probit

Figure 8. Estimated Root Mean Square Prediction Errors in Model 3.



Nonparametric



Probit

Figure 9. Estimated Relative Prediction Efficiency in Model 1.

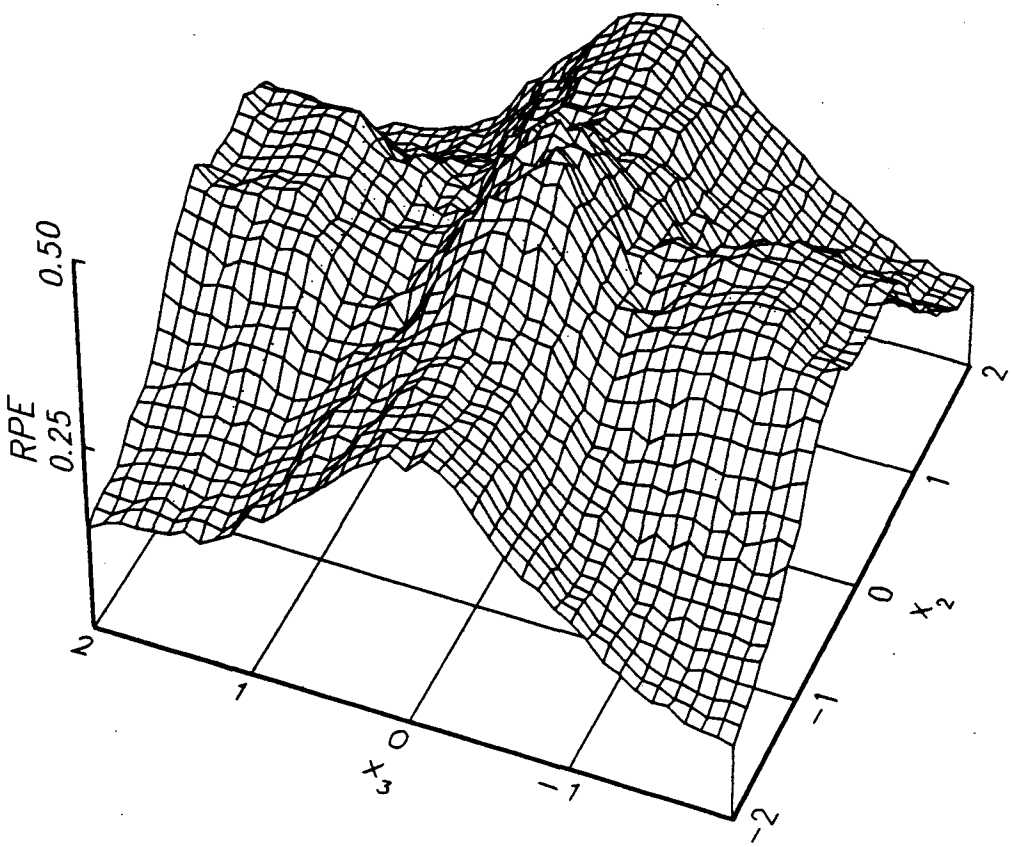


Figure 10. Estimated Relative Prediction Efficiency in Model 2.

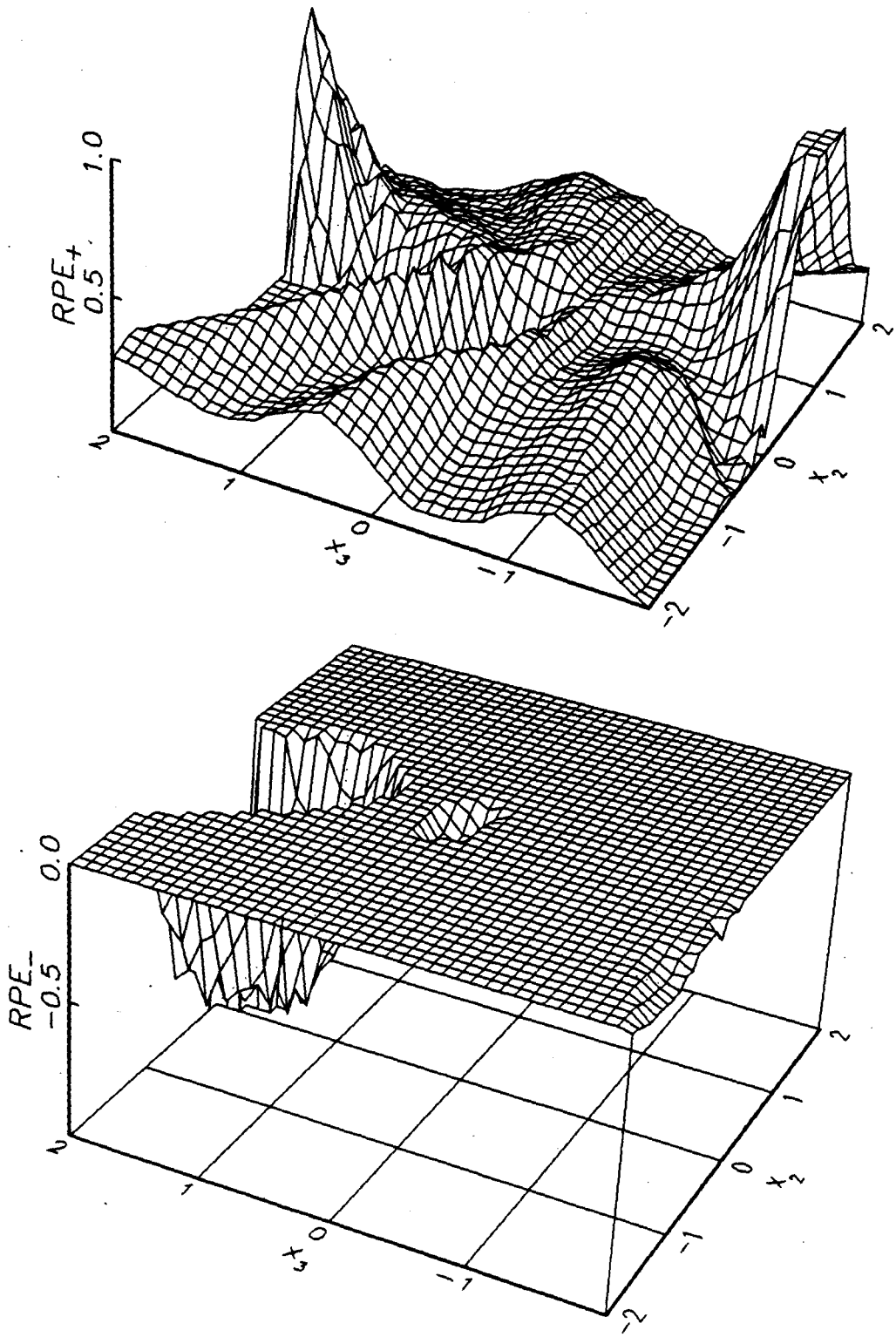


Figure 11. Estimated Relative Prediction Efficiency in Model 3.