Predicting Binary Outcomes*

Graham Elliott
Department of Economics
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0508

Robert P. Lieli
Department of Economics, C3100
University of Texas, Austin
1 University Station
Austin, TX 78712

June 2010

Abstract

We address the issue of using a set of covariates to categorize or predict a binary outcome. This is a common problem in many disciplines including economics. In the context of a prespecified utility (or cost) function we examine the construction of forecasts suggesting an extension of the Manski (1975, 1985) maximum score approach. We provide analytical properties of the method and compare it to more common approaches such as forecasts or classifications based on conditional probability models. Large gains over existing methods can be attained when models are misspecified. The results are informative for both forecasting environments as well as program allocation where the value of including the participant in the program depends on how useful the program turns out to be for that participant.

*Graham Elliott is grateful to the NSF for financial assistance under grant SES 0111238. An early paper on the subject was part of Robert Lieli's dissertation at the University of California, San Diego, completed in 2004. We have benefitted from discussions with Halbert White, Gary Chamberlain, Stephen Donald, Patrick Fitzsimmons, Bo Honoré, Mark Machina, Augusto Nieto Barthaburu, Max Stinchcombe and Imre Tuba. All errors are our responsibility.

1 Introduction

Constructing empirical models for the forecasting of binary outcomes and making binary decisions are problems that arise often in economics as well as other sciences. Examples in forecasting include predicting firm solvency, the legitimacy of credit card transactions, directional forecasts of financial prices, whether a loan is paid off or not, whether an introduced foreign plant species will become invasive or not. Such forecasts are often translated into decisions which are binary in character — the loan is granted or it is not, the student is admitted to the school or not, the candidate is hired or not hired, the surgery is undertaken or it is not, importation of a foreign plant species is allowed or not. Various statistical approaches to binary classification are available in the literature, from discriminant analysis, logit or probit models to less parametric estimates of the conditional probability model for the outcome variable (which is then employed to make the decision).

Typically, most estimation techniques used for binary classification do not make use of the loss function implicit in the underlying decision/prediction problem. For example logit and probit models are estimated to maximize the likelihood of the model, irrespective of the relative usefulness of true positives or true negatives. Nonparametric methods seek the best fit for the conditional probability based on the estimation technique loss function (typically squared error) rather than the appropriate loss function for the decision problem. In most applications, the relative costs of making errors—false negatives and false positives—are rarely balanced in the way that could be used to motivate these approaches. In detecting credit card fraud, "wasting" resources on checking that the customer has control over their credit card is perhaps less costly than failing to do so when their credit card number has been stolen. The main motivation for considering the approach suggested here is that even with local misspecifications that are difficult to detect using standard specification tests, parametric models of the conditional probability of a positive outcome can perform arbitrarily poorly when the loss function is ignored at the estimation stage. The semiparametric approach described below however requires far less information for the method to attain maximal utility, and through the utilization of the loss function at the estimation stage has useful properties given any misspecification.

We derive from a utility framework in a general setting the nature of optimal rules and draw out the statistical implications of these rules. A number of theoretical results follow. First, knowledge of the conditional probability of a positive outcome is shown to be sufficient for optimal prediction but not necessary. Optimal rules do not pin down a unique function to be estimated, however do describe features of the optimal functions. This gives rise to additional flexibility in modelling and suggests directly estimating the points at which the decisions switch from a positive to negative outcome or vice versa. Additional flexibility in models that do not translate to allowing the function to change sign will not allow a better fit, simplifying model selection. Rather than requiring the true specification of the conditional probability of a positive outcome to be known, models that are sufficiently flexible to arrive at the same set of decisions are suggested. This latter class of models is wider and hence easier to correctly specify. Since these decisions depend on the form of the utility function, the importance of taking into account parameters of the utility function in estimating the optimal function is shown.

The estimation method suggested employs the sample analog of the expected loss and builds upon and extends results of Manski (1985) and Manski and Thompson (1989). The results of Manski (1985) are extended to the more general optimand and to allow for serial correlation as is typically found in forecasting environments. We also provide asymptotic normality results for estimated utility. These results both establish the correct rate of convergence for the utility attained but also allow confidence intervals to be constructed for the estimated average utility. The result allows for pairwise testing of competing model specifications. The methods are analyzed in a Monte Carlo setting.

The next section describes the utility setup and the optimal forecasting/classification problem. It is there that the main insights as to what is important in this problem are gained. Section 3 examines the estimation approach we are proposing, establishing analytic results that suggest it will have reasonable properties in practice. In Section 4 numerical work shows that the proposed method can indeed lead to improvements in classification (relative to logit models and some off-the-shelf nonparametric methods) in practically relevant situations. Section 5 concludes.

2 The Forecasting Framework and General Results

We are interested in making a binary decision or forecast that can be characterized as setting action a to either one or minus one for the two possible decisions respectively. Hence we could assign $a = 1$ to be the decision to make a loan, or to go long in a particular security. The outcome is some binary random variable Y with sample space $\{1, -1\}$. For example, if the loan is paid back we set $Y = 1$ and if not $Y = -1$. We do not observe Y at the time the decision is made, hence the decision maker must predict or forecast this outcome based on a number of observables. These observed data for each individual or date are denoted by the k-dimensional vector X. The utility function of the decision maker depends on both the action and the outcome of the variable to be forecast, as well as potentially all or some subset of the observed covariates X, denoted $U(a, Y, X)$. Since the action and outcome are both binary, we have for any X just four possibilities. These can be described as

$$U(a, y, x) = \left\{ \begin{array}{cc} u_{1,1}(x) & \text{if } a = 1 \text{ and } y = 1 \\ u_{1,-1}(x) & \text{if } a = 1 \text{ and } y = -1 \\ u_{-1,1}(x) & \text{if } a = -1 \text{ and } y = 1 \\ u_{-1,-1}(x) & \text{if } a = -1 \text{ and } y = -1 \end{array} \right\}$$

A number of problems fit this framework. The decision to extend a loan to an applicant under the uncertainty over whether or not they will repay the loan we have mentioned above. In such situations, it may well be that the utility function depends directly on some of the aspects of the individual seeking the loan. For example the head of the International Finance Corporation, "the private sector arm of the World Bank", discusses their dilemma of how to balance the conflict between making loans that are profitable and at the same time contribute to the development of certain target groups (regions, industries etc.)[1]. Here the value of a successful loan to the IFC depends on how needy the recipient was in the first place, which no doubt affects the chances of being repaid. In the example concerning the import of foreign plants, a number of biological traits can be used to predict invasiveness. If the species does become invasive, these same traits will often be related to the amount of damage caused and the cost of control (Lieli and Springborn (2010)). In both of these

[1] "Moderniser tackles the taboos", Financial Times, November 4, 2003.

examples the role of the X covariates is twofold — they are informative about the outcome and but also directly enter the decision maker's utility function.

The following condition is not restrictive and gives content to the problem.

Condition 1 *(Utility function)*

(i) $u_{1,1}(x) > u_{-1,1}(x)$ *and* $u_{-1,-1}(x) > u_{1,-1}(x)$ *for all* x *in the support of* X*;*

(ii) $u_{k,l}$ *is Borel measurable and bounded as a function of* x*, i.e.* $|u_{k,l}(x)| < M_1$ *for some* $M_1 > 0$*, all* x *and* $k, l \in \{-1, 1\}$.

Part (i) requires that the utility gained from matching the correct action to the correct outcome results in higher utility than an incorrect matching, for any possible value of the covariates X, i.e. making no error is better than making an error, a typical property of loss functions (see Granger 1969). The assumption of bounded utility in part (ii) should not limit the scope of practical applications, it ensures that expected values of quantities used below actually exist and simplifies some technical arguments.

Since Y is not observable at the time of the decision the decision maker will maximize expected utility conditional on the observed data $X = x$, i.e. the decision maker chooses the action to solve the maximization problem

$$\max_a E[U(a, Y, X) \mid X = x]. \tag{1}$$

Denote the conditional probability that $Y = 1$ given $X = x$ as[2] $p(x)$. When $p(x)$ is known the optimal decision follows from integrating out the unknown value for Y. The optimal action is to choose $a = 1$ if the conditional probability exceeds a cutoff that depends on the utility function, i.e. choose $a = 1$ if and only if

$$p(x) > c(x) \equiv \frac{u_{-1,-1}(x) - u_{1,-1}(x)}{u_{1,1}(x) - u_{-1,1}(x) + u_{-1,-1}(x) - u_{1,-1}(x)}. \tag{2}$$

The interpretation of this result follows from noting that $u_{1,1}(x) - u_{-1,1}(x)$ is the gain from getting the decision correct when $Y = 1$ and $u_{-1,-1}(x) - u_{1,-1}(x)$ is the gain from getting the decision right when $Y = -1$. The cutoff $c(x)$ is higher the greater the relative

[2] Here we explictly restrict $p(x)$ to be independent of the forecast a. This rules out feedback from the action (allocation) to the outcome of interest.

gain in getting the decision correct when $Y = -1$ compared to when $Y = 1$. We thus choose a higher cutoff to more often choose $a = -1$ when the gain from being correct in this case is larger. Condition 1 ensures $c(x)$ is strictly between zero and one for any x. This calculation, when the utility function does not depend on X, has been made in many previous papers, see Boyes et. al. (1989), Granger and Pesaran (2000), Pesaran and Skouras (2002).

The optimal problem (1) can equivalently be written as

$$\max_{a(\cdot)} E_{Y,X}\{U\left[a(X), Y, X\right]\}, \tag{3}$$

where the maximization is undertaken over the space of measurable functions with range $\{-1, 1\}$ defined on $\mathbb{R}^k$. The resulting optimal action or forecast $a^*(X)$ partitions the support of X into two parts, that which corresponds to a positive action and that for a negative action. We can represent the set of possible predictors of Y as $a(X) = sign[g(X)]$, $g \in G$ where G is the set of all measurable functions from $\mathbb{R}^k$ to $\mathbb{R}$ (we define $sign(z) = 1$ for $z > 0$ and $sign(z) = -1$ for $z \leq 0$). We can rewrite the problem (3) as

$$\max_{g \in G} E_{Y,X}\left\{b(X)[Y + 1 - 2c(X)]sign[g(X)]\right\}, \tag{4}$$

where $b(x) = u_{1,1}(x) - u_{-1,1}(x) + u_{-1,-1}(x) - u_{1,-1}(x) > 0$.[3] This function presents the range of loss functions for binary prediction problems that correspond to maximizing utility. This result, in the special case of $b(x)$ and $c(x)$ constant, is equivalent to one of the cases in Manski and Thompson (1989), where loss is L_1 in their notation, their parameter $a = 1 - c$ and their function $\theta(x)$ is equivalent to $g(X) + c$ in our notation.

A primary insight of the paper is that the optimizing function is not unique. Let G^* be the set of all (measurable) functions on $\mathbb{R}^k$ whose *sign* produces the same optimal partition of the covariate space as the solution of $a^*(x)$ of the problem (3). Hence, we can write $a^*(X) = sign[g^*(X)]$ for $g^* \in G^*$. Equation (2) suggests choosing $g^*(x) = p(x) - c(x)$. However for an optimal forecast full knowledge of $p(x)$ is sufficient but not necessary. Any function $g^* \in G^*$ can be considered to be a correctly specified model for the purposes of prediction. This can be most easily seen from Figure 1. The optimal forecast simply

[3]Rearranging, $U(a, y, x) = \frac{1}{4}b(x)[y + 1 - 2c(x)]a + \frac{1}{4}b(x)[y + 1 - 2c(x)] + u_{-1,y}(x)$. Dropping terms independent of a, substituting $a(x) = sign[g(x)]$ and taking expectations yields (4).

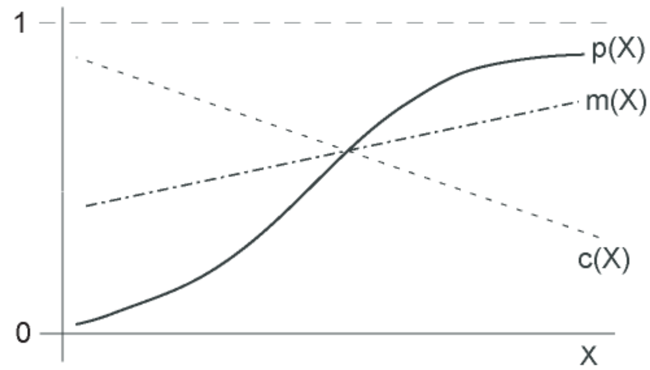

Figure 1: Here $p(X)$ gives the probability that $Y = 1$ given scalar X, $c(X)$ gives the cutoff for the decision rule, $m(X)$ gives a function that differs from $p(X)$ everywhere but at the cutoff and so delivers the same decisions.

involves knowing which side $p(x)$ is of $c(x)$. Consider the function $\tilde{g}(x) = m(x) - c(x)$ such that $sign[\tilde{g}(x)] = sign[p(x) - c(x)]$. The function $m(x)$ differs from $p(x) - c(x)$ almost everywhere in x—everywhere except that they are equal at the points where $p(x)$ cuts $c(x)$ and is always above $c(x)$ when $p(x)$ is above $c(x)$. As such, the forecasts and hence utility that result from using $\tilde{g}(x)$ are identical to those constructed using $p(x) - c(x)$.

This insight has extremely useful practical implications for both estimation and model specification. Manski and Thompson (1989) note that for their linear model and $p(x)$ either monotone or isotonic then there is a single crossing allowing the model to be well specified only with this information. For this model they apply the Manski (1985) maximum score principle that bypasses the estimation of $p(x)$. In the next section we will employ this insight extending the Manski (1985) maximum score principle to allow for a wider array of (possibly misspecified) forecast models, problems with more than a single crossing point, and the more general loss function. The method will provide asymptotically the best crossing points, or at least the most valuable ones as measured by utility. We also examine model selection procedures, noting that good model selection requires flexible enough functions of the covariates so as to capture the possibility of multiple crossings between the cutoff and conditional probability rather than a good approximation of $p(x)$ itself.

Our result also shows the role of the loss function in determining optimal rules. It is in the roles of $b(x)$ and $c(x)$ in reweighting utility (discussed in context with estimation in the next section) along with the insight that only crossing points need to be estimated well that problems with the commonly applied two step approach to binary prediction of estimating $p(x)$ without reference to the loss function arise. Typically $p(x)$ is estimated with reference to an arbitrary loss function. For example probit and logit are estimated using maximum likelihood (ML), an approach which is designed to provide the best global fit for the model of $p(x)$ to the data independent of both where it crosses $c(x)$ and the relative importance of various regions in X of having the correct forecast. Other methods such as neural networks use quadratic loss, which suffers from the same problems as ML. We show in Section 4 that these methods can be very poor predictors, even when they are close to being correctly specified models.

3 Estimating Binary Prediction Models

3.1 The Maximum Utility (MU) Estimator

The optimal forecast/allocation method chooses a function $g^*(\cdot)$ that solves (4). For a solution to this optimization problem, the forecaster must search over a function space. More reasonably in practical situations the set of possible functions will be restricted and a constrained optimization will be undertaken. Instead of considering all the possible functions we restrict ourselves to a subset H of G and work with predictors of the form

$$P(H) = \{sign[h(X)] : h \in H\}.$$

We will parameterize the elements of H as $h(x) = h(x, \theta)$,where $\theta \in \Theta \subseteq \mathbb{R}^p$, $p \in \mathbb{N}$; we will write H_Θ for the parameterized class. Hence we have a parametric model for the predictor function that is known up to the $p-$dimensional vector of unknown coefficients θ. The optimal predictor is then obtained by solving

$$\max_{\theta\in\Theta} S(\theta) \equiv \max_{\theta\in\Theta} E_{Y,X}\{b(X)[Y + 1 - 2c(X)]sign[h(X, \theta)]\}.$$

There is of course a cost to restricting the form of the predictors considered. It may well be that the set of functions that maximize expected utility in the unconstrained problem, G^*, are 'locked out' of H_Θ, meaning that the set of optimal functions in the constrained problem, H_Θ^*, is also mutually exclusive of G^*, Nevertheless if

$$\exists\theta^* \in \Theta \text{ such that } sign[h(x,\theta^*)] = sign[p(x) - c(x)] \text{ for all } x \in \text{support}(X), \tag{5}$$

then $G^* \cap H_\Theta^*$ is nonempty. In standard econometric language this means that for the model to be optimal from a forecasting or classification standpoint, it does not have to be fully correctly specified for $p(x) - c(x)$; it is enough for it to be correctly specified for the sign of $p(x) - c(x)$. This gives clear insights in how to choose $h(X,\theta)$ (up to knowing θ) in practice. Essentially all that is required to achieve the same prediction as knowing $p(x)$ is that the model be sufficiently flexible to capture all of the crossing points of $p(x)$ and $c(x)$.

The weaker requirement than correct specification of $p(x)$ can be exploited practically in simplifying model building. For example if we have a single predictor ($k = 1$) then a polynomial in X of order d allows for d changes in the sign of the predictor as a function of x. Knowing $p(x) - c(x)$ only up to the number of sign changes is thus sufficient for obtaining a correct specification of the forecasting model, even though $p(x)$ itself is not known well enough for correct specification. Knowing $p(x)$ is monotonically increasing in x (and knowledge of $c(x)$ which comes from knowing the utility function) tells us exactly the number of crossing points between $p(x)$ and $c(x)$ but does not tell us the correct parameterization of $p(x)$. For higher dimensional problems ($k > 1$) we can still use this insight, however concrete implications become more difficult. Results for interesting cases are available however, as shown in the following proposition.

Proposition 1 *Let $\tilde{c} \in \mathbb{R}, \beta \in \mathbb{R}^k, \beta \neq 0$, and let $Q : \mathbb{R} \to \mathbb{R}$ be a function such that the set $\Sigma = \{\sigma : Q(\cdot) - \tilde{c}$ changes sign at $\sigma\}$ has at most a finite number of elements. Then there exists a finite order polynomial $h(x)$ and $A \subset \mathbb{R}^k$ with Lebesgue measure one such that $sign[h(x)] = sign[Q(x'\beta) - \tilde{c}]$ for all $x \in A$. If in addition Q is lower semicontinuous at each $\sigma \in \Sigma$, then $A = \mathbb{R}$.*

Some examples can clarify the result. If $c(X)$ is constant, $Y^* = Q(X'\beta) + U, Y = sign[Y^*]$, and U is independent of X with cdf $F_U(u)$, then $p(X) = 1 - F_U[-Q(X'\beta)] > c$ if and only if $Q(X'\beta) > \tilde{c}$, for $\tilde{c} = -\inf\{u : F_U(u) \geq 1 - c\}$ (see for example Horowitz (1998), Powell

(1994)). For a second example suppose that U is only $(1-c)-$quantile independent of X, i.e. $\inf\{u : F_{U|X}(u) \geq 1-c\}$ is a constant $-\tilde{c}$ that only depends on c but not the value of X. Then it remains true that $p(X) > c$ if and only if $Q(X'\beta) > \tilde{c}$. This example provides a generalization of the 'single crossing' restriction of Manski and Thompson (1989).

If the set of maximizers of $S(\theta)$ is nonempty, then it will often contain more than one element. We distinguish between two types of multiplicity. On the one hand,

$$P\{sign[h(X,\theta_1)] = sign[h(X,\theta_2)]\} = 1 \tag{6}$$

may hold for distinct θ_1, θ_2 in $\arg\max S(\theta)$, i.e. the same optimal partition of the covariate space under the model H_Θ may be induced by more than one value of θ. Alternatively, it might happen that the optimal partition itself is not unique so that θ_1 and θ_2 both maximize $S(\theta)$, but (6) fails to hold.

Multiple maxima of the first type arise quite generally; an example is that the linear model is only identified up to scale (Manski (1985), Horowitz (1992)). Another example is provided by discrete X — in this case most values of θ will have an entire neighborhood of points leading to the same decision rule. In contrast, multiple maxima of the second type are non-generic. Their existence is predicated on rare coincidences between the decision maker's utility function, the parameterization H_Θ, and the joint distribution of Y and X.

While lack of identification is an issue when θ is a structural parameter endowed with economic content, multiple maxima of $S(\theta)$ are of little concern in the present forecasting or classification context. First, the model H_Θ is not intended to play a role or have any meaning beyond approximating the optimal decision rule $sign[p(x) - c(x)]$. Second, taking the utility maximization problem literally means that the decision maker will be indifferent between two parameter vectors that result in the same level of utility. For these reasons our analysis will focus on the properties of the optimand function instead of the usual econometric focus on the optimizers, which could be considered as nuisance parameters of the forecasting problem.

To estimate a member of H^*_Θ based on a finite set of data, we suggest the sample analog of the expected utility maximization problem that results in the set of models H_Θ. Given a sample of observations $\{(Y_i, X_i')\}_{i=1}^n$, choose θ to solve

$$\max_{\theta\in\Theta} S_n(\theta) \equiv \max_{\theta\in\Theta} n^{-1}\sum_{i=1}^{n} b(X_i)[Y_i + 1 - 2c(X_i)]sign[h(X_i, \theta)].$$

For $b(X_i)$ constant, $c(X_i) = 0.5$ and $h(X_i, \theta) = X_i'\theta$ this is Manski (1985)'s maximum score estimator. When the estimation procedure searches over candidate functions, the goal is to match the sign of Y_i with the sign of $h(X_i, \theta)$. By determining the value of each match through the weights $b(X_i)[Y_i + 1 - 2c(X_i)]$, the utility function plays a direct role in the estimation of the parameters. As earlier, $c(X_i)$ determines the relative value of correct prediction when $Y_i = 1$ relative to correct prediction when $Y_i = -1$, whereas $b(X_i)$ controls the value of correctly predicting Y_i for that individual regardless of Y_i. This term is only important if the model is not correctly specified, i.e. the model H_Θ cannot possibly reproduce the theoretically optimal predictor for all values of X. In this case the trade-offs in finding the expected utility maximizing fit conditional on model specification will be guided partly by the function $b(.)$. If the model were correctly specified one could replace $b(.)$ by any other bounded positive function (which may not correspond to the utility function at all) and the fitted model will still reproduce $sign[p(x) - c(x)]$ asymptotically. (Small sample fit would be affected).

Given the observed data, $S_n(\theta)$ can take on at most 2^n different values as a function of θ, regardless of the specification of $h(x, \theta)$. For each $\theta, S_n(\theta)$ is a sum of n terms, only the sign of each term depends on θ. Thus even if each sign could be set independently of the others, the sum could take on 2^n distinct values. As the range of $S_n(\theta)$ is finite over Θ, a maximum must always exist.

It is clear that maximization of $S_n(\theta)$ in practice cannot be undertaken by hill climbing methods. The Monte Carlo studies presented in Section 4 demonstrate that the simulated annealing algorithm (see Corona et. al. (1987) and Goffe et. al. (1984)) is robust enough to handle the nonstandard feature of the objective function under consideration here.

3.2 Asymptotic Properties of the MU estimator

As noted above, instead of studying the convergence of $\hat{\theta}_n$, we will focus on the convergence properties of the maximized value of the sample objective. We first establish uniform convergence of $S_n(\theta)$ to $S(\theta)$, consistency follows directly. Under additional assumptions, $\max_\theta S_n(\theta)$ converges at the usual rate and is asymptotically normal.

Using only the assumption that the data are i.i.d., Manski (1985, Lemma 4) establishes a.s. uniform convergence of $S_n(\theta)$ to $S(\theta)$ in the special case when $h(x,\theta) = x'\theta$, and the utility function does not depend on x. The proof cites a Glivenko-Cantelli type result by Rao (1962) whose applicability depends critically on the geometry of the linear model. As we allow for general functional forms and serially dependent data, we need to extend Manski's result. We impose some (joint) restrictions on the model H_Θ and the distribution of X.

Condition 2 *(i) $\Theta \subset \mathbb{R}^p$ is compact; (ii) The function $(x,\theta) \mapsto h(x,\theta)$ is jointly Borel measurable and Lipshitz-continuous over Θ, i.e. there exist constants $M_2 > 0$ and $\lambda > 0$ such that $|h(x,\theta) - h(x,\theta')| \leq M_2 \|\theta - \theta'\|^\lambda$ for all $\theta, \theta' \in \Theta$ and $x \in$ support(X).*

As θ varies over Θ, $h(X,\theta)$ might act as a continuous variable for some values of θ and as a discrete one for others. In the latter case it is possible that $h(X,\theta) = 0$ happens with positive probability, creating a discontinuity in the function $S(\theta)$. The following condition ensures these discontinuities do not interfere with uniform convergence.

Condition 3 *There exists a subset Θ_0 of the parameter space such that (i) $\sup_{\theta\in\Theta_0} P[-\delta \leq h(X\ ,\theta) \leq \delta] \leq M_3\delta^r$ for some constants $M_3 > 0, r > 0$ and $\forall\delta > 0$ sufficiently small; (ii) if $\theta \notin \Theta_0$, then the function $x \mapsto h(x,\theta)$ is either constant, or depends at most on discrete, finitely supported components of X.*

Restricted to Θ_0, $S(\theta)$ is a (uniformly) continuous function of θ, while restricted to Θ/Θ_0 it is a step function (Lemma 2 in appendix). It is possible for Θ_0 to be equal to the entire parameter space or be empty. Condition 3(i) is somewhat stronger than having $P[h(X,\theta)] = 0] = 0$ for all $\theta \in \Theta_0$. It is satisfied if for each θ in Θ_0 the distribution of the random variable $h(X,\theta)$ is absolutely continuous (possibly conditional on some components of X), and the associated densities are bounded in a neighborhood of zero uniformly over Θ_0. For example, let $h(X, \theta) = \theta_1 + \theta_2 X$, $\theta \in [-1, 1] \times [-1, 1]$, where X is has bounded density $f(x)$ with $xf(x)$ also bounded. Then $h(X, \theta)$ has the density $z \mapsto \frac{1}{\theta_2} f\left(\frac{z-\theta_1}{\theta_2}\right)$ for $\theta_2 \neq 0$. This density is bounded for, say, $z \in (-0.1, 0.1)$ uniformly over $\theta \in \Theta_0 = \{\frac{1}{2}\} \times ([-1, 1]\backslash\{0\})$, and for $\theta \notin \Theta_0$, $h(X, \theta)$ is constant as a function of X. In contrast, Manski (1985) and Horowitz (1992) normalize the parameter space (and restrict the distribution of X) to ensure $P(X'\theta = 0) = 0$ for all $\theta \in \Theta$, essentially requiring $\Theta = \Theta_0$. In a pure forecasting context such normalizations would be hard to motivate — Condition 3 makes them unnecessary.

Finally, we restrict the sampling process, while allowing for serially dependent observations recorded over time.

Condition 4 *For p, λ and r as in Conditions 2 and 3, let W denote the smallest even integer greater than $4p/r\lambda$. Then $\{(Y_i, X_i')\}_{i=1}^n$ is a strictly stationary, mixing sequence of observations from (Y, X'), defined on a complete probability space, with strong mixing coefficients $\alpha(d) = O(d^{-A})$, where $A > (W-1)(1+W)/2$.*

The following result extends Manski (1985, Lemma 4).

Proposition 2 *Given Conditions 1 through 4, then (a) $\sup_{\theta\in\Theta} |S_n(\theta) - S(\theta)| \to 0$ w.p.1 as $n \to \infty$, (b) $\max_{\theta\in\Theta} S_n(\theta) \to \sup_{\theta\in\Theta} S(\theta)$ w.p.1 as $n \to \infty$. Further if $\hat{\theta}_n$ is a measurable selection from $\arg\max_{\theta\in\Theta} S_n(\theta)$, then (c) $S(\hat{\theta}_n) \to \sup_{\theta\in\Theta} S(\theta)$ w.p.1 as $n \to \infty$.*

It is part (c) of Proposition 2 that provides the basic justification of the proposed estimation method. It states that the expected utility derived from classifications using $\hat{\theta}_n$ approaches the upper bound of utility levels achievable under the parameterization H_Θ, regardless of whether this parameterization satisfies (5). The result relies on the function we are maximizing and the target function getting close asymptotically, but does not rely on the maximizer $\hat{\theta}_n$ getting close to a 'true' set of optimal coefficients. In fact, the conditions of Proposition 2 are not generally sufficient to guarantee that there exists an optimal value of θ that attains the supremum. In contrast to the above, the existence of a 'nice' set of maximizers of $S(\theta)$ is assumed to show asymptotic normality of $\max_\theta S_n(\theta)$. No analogous result is given in the maximum score literature.

Condition 5 *Let* $\Theta^* = \arg\max_{\theta\in\Theta} S(\theta)$ *and* $\Theta^*_\varepsilon = \{\theta \in \Theta : d(\theta, \Theta^*) \leq \varepsilon\}$. *(i)* $\Theta^* \neq \emptyset$ *and* $\Theta^*_\varepsilon \subset \Theta_0$ *for some* $\varepsilon > 0$. *(ii)* $\sup_{\theta\in\Theta\backslash\Theta^*_\varepsilon} S(\theta) < \max_{\theta\in\Theta} S(\theta)$ *for all* $\varepsilon > 0$. *(iii)* $\theta, \theta \in \Theta^*$ *implies (6). (iv)* $P[sign(h(X, \theta^*)) = Y] \neq 1$ *for* $\theta^* \in \Theta^*$.

Whilst these are fairly high level conditions, they are not unreasonable. Given that $S(\theta)$ is continuous (and bounded) over Θ_0, and is a step function otherwise, the existence of a maximum is not a big leap of faith — the only part of the parameter space that can potentially cause problems, when $\Theta \neq \Theta_0$ and $\Theta_0 \neq \emptyset$, is the boundary of Θ_0. In addition, Condition 5(i) requires that $h(X, \theta)$ acts as an absolutely continuous random variable for values of θ in the neighborhood of Θ^*. Barring unusual parameterizations, this will be the case if, conditional on the rest of X, some component of X is absolutely continuous and provides some useful information about Y. Neither this condition nor Condition 3 imposes full support or smoothness on the conditional density (or a related density), unlike Manski (1985), Horowitz (1992) and Kim and Pollard (1990). Condition 5(ii) states that the function $S(\theta)$ has a 'clean' maximum, i.e. one cannot get arbitrarily close to the optimum in a part of the parameter space that is disconnected from Θ^*. Condition 5(iii) rules out non-generic type of multiple maxima discussed in Section 3.1, Condition 5(iv) disallows perfect predictors.

We now have a set of conditions sufficiently strong to establish the asymptotic distribution of $\max_\theta S_n(\theta)$. Define the asymptotic variance function as

$$V(\theta) = var[s(Y_1, X_1, \theta)] + 2\sum_{m=2}^{\infty} cov[s(Y_1, X_1, \theta), s(Y_m, X_m, \theta)],$$

where $s(y, x, \theta) = b(x)[y + 1 - 2c(x)]sign[h(x, \theta)]$. For i.i.d. data the covariance terms vanish and the formula reduces to

$$V(\theta) = E\left\{b(X)^2[Y + 1 - 2c(X)]^2\right\} - [S(\theta)]^2.$$

Hence, in this case $V(\theta)$ depends on θ only through $S(\theta)$, and is therefore constant over Θ^*. For serially dependent data the covariance terms can be written as

$$E\Big\{b(X_1)[Y_1 + 1 - 2c(X_1)]sign[h(X_1, \theta)]b(X_m)[Y_m + 1 - 2c(X_m)]sign[h(X_m, \theta)]\Big\} - [S(\theta)]^2.$$

Under Condition 5(iii), $sign[h(X, \theta_1^*)] = sign[h(X, \theta_2^*)]$ w.p. 1 for any two maximizers $\theta_1^*, \theta_2^* \in \Theta^*$. Therefore, given this additional assumption, $V(\theta^*)$ does not vary with $\theta^* \in \Theta^*$ even when the data points are serially dependent. Let S^* and V^* denote the values of $S(.)$ and $V(.)$ over Θ^*. Then:

Proposition 3 *Given Conditions 1-5, (a)* $n^{1/2}[\max_{\theta\in\Theta} S_n(\theta) - S^*] \longrightarrow^d N(0, V^*)$, $V^* > 0$, *and (b)* $n^{1/2}[S(\hat{\theta}_n) - S^*] = o_p(1)$.

This result establishes the rate of convergence of the decision maker's utility, both in sample and out of sample. For in sample evaluation (evaluation of $\max_{\theta\in\Theta} S_n(\theta)$), the convergence rate is exact. When a sample estimate is used for out of sample evaluation without re-estimation of the parameter (evaluation of $S(\hat{\theta}_n)$ where both the estimation sample and evaluation sample become large), the convergence rate is a lower bound. With an estimate of V^*, part (a) also allows construction of a confidence interval for S^*, the maximum utility achievable under the model specification. This result can also be used to test that utility achieves a certain level.

In the case where the data is i.i.d. Proposition 3 can be employed to test between specifications[4]. Suppose that we have a model $h_1(x, \theta_1)$ and wish to compare with another model $h_2(x, \theta_2)$, where $h_2(x, \theta_2)$ nests $h_1(x, \theta_1)$. For example, $h_1(x, \theta_1)$ could be a linear specification, while $h_2(x, \theta_2)$ could also include quadratic terms. We can test the null hypothesis that $S_1(\theta_1^*) = S_2(\theta_2^*)$ against the alternative that $S_1(\theta_1^*) < S_2(\theta_2^*)$. Construct an i.i.d. sequence of Bernoulli(λ) random variables $\{Z_i\}_{i=1}^n$, independent of the data, and compute

$$\hat{S}_{1n} = \max_{\theta_1\in\Theta_1} S_{1n}(\theta_1) \equiv \max_{\theta_1\in\Theta_1} n^{-1}\sum_{i=1}^n b(X_i)[Y_i + 1 - 2c(X_i)]sign[h_1(X_i, \theta_1)]$$

and

$$\hat{S}_{2n}^\lambda = \max_{\theta_2\in\Theta_2} S_{2n}^\lambda(\theta_2) \equiv \max_{\theta_2\in\Theta_2} n^{-1}\sum_{i=1}^n Z_i b(X_i)[Y_i + 1 - 2c(X_i)]sign[h_2(X_i, \theta_2)].$$

Hence, one specification is estimated on the full sample, while the other on a random subsample picked out by the Bernoulli variables. It is straightforward to generalize the proof of Proposition 3 to show that the asymptotic distribution of $\hat{S}_{1n}$ and $\hat{S}_{2n}^\lambda$ is jointly normal and

[4] There are a large number of methods for model selection for classification schemes, although none have been shown to extend to the general methods of this paper.

derive an explicit expression for the variance-covariance matrix (see the Appendix). Under the the null hypothesis it follows in particular that

$$n^{1/2}\left(\lambda\hat{S}_{1n}-\hat{S}_{2n}^{\lambda}\right)\longrightarrow^{d}N(0,\sigma_{\lambda}^{2}),$$

where $\sigma_\lambda^2=\lambda(1-\lambda)E[b(X)^2(Y+1-2c(X))^2]$. Using the sample analog principle, a consistent estimator for σ_λ^2 is easy to construct.

4 Finite Sample Properties: Monte Carlo Results

We present simulation evidence to give both empirical content to the theoretical results in Section 2 that suggest potential issues in estimating $p(x)$ to construct forecasts and also to demonstrate the performance of the methods suggested in Section 3. To do this we present two data generating processes (DGP's) and two sets of utility functions for each DGP. As utility has no natural units, we will compare the ratio of utility for any model to the utility that would be gained if $p(x)$ were known (the PK model, where infeasible forecasts are constructed from the known $p(x)-c(x)$). We will refer to this ratio as relative utility (RU)[5].

For each DGP and utility function we present both in and out of sample results for the prediction method. For the same reasons that R^2 is always nondecreasing in the number of terms in a linear regression, by the properties of maximization the in sample results will always favor larger models. By the same token, they will often outperform the PK model for the sample. The in sample results are none-the-less interesting as a gauge of the extent of the 'overfitting' of the methods to the data. The larger the gain over the PK model the greater the ability of the model and method to find incidental matches in the data that do not reflect features of the DGP. The out of sample results are of course directly interesting as they examine how well the estimation fits the model on a new set of data.

The first model for the Monte Carlo study is given by $p(X)=\Lambda(-0.5X+0.2X^3)$ where $\Lambda(.)$ is the logit function. The distribution for X is $U(-2.5,2.5)$. Figure 2 shows $p(x)$ over this support. We draw 75 in sample observations (for estimation) and 5000 out of sample

[5] Since a constant added to $S(\cdot)$ will change the ratio, we normalize as follows. Set $u_{-1,-1}(x)=u_{-1,1}(x)=0$, which amounts to representing expected utility as $0.25S+0.25E\left[b(X)(Y+1-2c(X)\right]$.

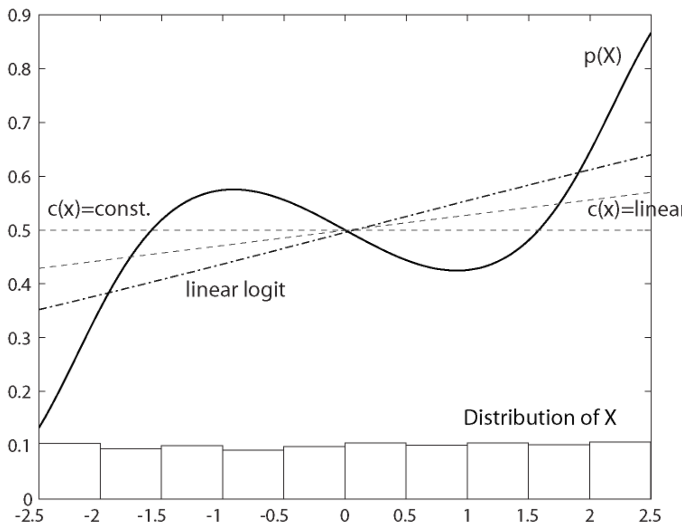

Figure 2: The solid line shows $p(x)$. Dashed lines are $c(x)$ for P1 (flat line at 0.5) and P2 (upward sloped line). The dot-dash line shows an estimated linear probit model fit of $p(x)$. The histogram at the bottom of the figure is for a draw of the X variable.

observations (for evaluation). The number of Monte Carlo replications is 500. This model is chosen so that it is sufficiently flexible to enable multiple crossings (so that the categorization is not simply a crossing point) but the cubic logit model is correctly specified. The model $p(X)$ is sufficiently close to the single threshold model $\Lambda(-0.5X)$ that 5% Lagrange Multiplier (LM) tests of this null model reject for the model that includes both X^2 and X^3 with roughly 33% power only. We choose a local misspecification because it is precisely these types of misspecifications that are unknown to the researcher. In terms of utility we examine two sets of preferences, first setting $c(x) = 0.5$ (P1) and in the second $c(x) = 0.5 + 0.025x$ (P2), shown in Figure 1 by the dashed lines. We set $b(x)$ to be constant, the value of this constant has no effect on the relative utility. For both utility functions there are three crossing points. Results are presented in Table 1.

The cubic ML model here is correctly specified for $p(x)$ — it follows directly that the difference between this model and PK is due only to estimation error of the model parameters; asymptotically it will obtain the highest possible utility. It is no surprise then that

Table 1: Model 1 - Possibly Correctly Specified MLE

	ML			MU			
	Lin	Cubic	Mix 1	Lin	Cubic	Mix 1	Mix 2
	In/Out	In/Out	In/Out	In/Out	In/Out	In/Out	In/Out
RU (P1)	86.2/37.0	139.4/73.1	120.2/52.7	146.1/50.3	226.0/67.1	179.4/57.3	201.2/61.5
RU (P2)	77.2/11.6	140.9/67.1	115.1/36.0	139.1/31.9	230.9/61.3	177.0/45.2	206.8/53.7
Correct(P2)	58/53	63/57	61/54	63/54	71/56	67/55	69/56

Note: The first two rows denote relative utility (RU), as defined in the first paragraph of Section 4 and expressed as a percentage, for P1 and P2. The last row is the percentage of correctly classified observations for P2. "In" refers to the in sample result, "Out" to the out of sample result.

this method performs best. As there are three crossing points for both preferences, the cubic MU method is also correctly specified. Hence differences between this model and PK also represent estimation error. The correctly specified ML method performs better than the MU method, reflecting the gain of parametric methods over semiparametric methods when the parametric specification is correct. This effect however is small. Comparing the (misspecified) linear models MU performs far better than ML. A best linear choice would be the approximation that maximizes utility. However, as noted at the end of Section 2, the ML method instead provides a best fit of $p(x)$ over all of the range of X whereas MU is designed to focus estimation on getting the points and signs correct around the points where $p(x)$ cuts $c(x)$. With a linear specification this means choosing the best single crossing point of $c(x)$. Gains from doing so in a misspecified model can be very large, as we see in the Monte Carlo results. These differences are not well reflected in the percentage of correctly classified observations. For example under P2 the linear MU model correctly classifies 54% of the observations out of sample, just one percentage point less than the cubic model, but the difference in the relative utilities is large. We can also see that the size of the effect depends on the utility function; in Table 1 the gain is larger for the second set of preferences despite the differences between the preference configurations being small. If we believe that models are approximations, using a method that is robust such as MU over a method that is not such as ML seems appropriate.

In sample overfitting can be examined by comparing the in sample and out of sample results. Both ML and MU have a strong tendency to overfit in sample, however the problem seems more severe for the MU method. This creates challenges for model selection. We examine a standard hypothesis test approach to selecting between the two models where for each draw of the pseudo variables we test the null of the linear model against the cubic using a size 5% LM test. We then apply both estimators to the model chosen (denoted ML Mix 1 and MU Mix 1, respectively). For the MU method we also use a size 5% test with $\lambda = 0.75$ based on the asymptotic results at the end of Section 3 (denoted MU Mix 2). The power of the LM test to reject the linear model is 33% as noted above. This is enough to partially offset the very poor performance of the ML method in this misspecified model for both sets of preferences. However, relative out of sample utility for ML Mix 1 is still smaller than for MU Mix 1 (52.7% vs. 57.3% for P1 and 36% vs. 45.2% for P2). This is because the extent to which ML outperforms MU in the correctly specified model is overpowered by the nontrivial chance of misspecification and the extent to which MU outperforms ML in this case. The mixture based on the asymptotic theory of this paper is better again—-in the out of sample results this method gains 61.5% of available utility for P1 and 53.7% for P2. This additional improvement over the performance of MU Mix 1 is due to the fact that the MU based test rejects the linear model with almost 66% probability (the test has power precisely against those types of misspecifications that matter for optimal classification).

The second DGP for the Monte Carlo study is a special case of the models in Proposition 1. The conditional probability is given by $p(X) = \Lambda\left(Q(1.5X_1 + 1.5X_2)\right)$ where

$$Q(\sigma) = \frac{1.5 - 0.1\sigma}{\exp(0.25\sigma + 0.1\sigma^2 - 0.4\sigma^3)}.$$

The distribution for (X_1, X_2) is uniform on $[-3.5, 3.5] \times [-3.5, 3.5]$. Figure 3 shows $\Lambda(Q(\sigma))$ over part of the range of $1.5X_1+1.5X_2$ (solid line). This model corresponds to a heteroskedastic linear index model $Y^* = 1.5 - 0.1\sigma + \epsilon \exp\left[0.25\sigma + 0.1\sigma^2 - 0.4\sigma^3\right]$ with $Y = sign(Y^*)$. Two variations on the utility model are examined to highlight the effect of the utility function on the performance of each of the methods. In the first specification (P3) we set $c(x) = 0.75$ and $b(x) = 20$. The second (P4) varies $b(x)$ by setting it to 60 for $-1.5 < x_1 + x_2 < 1.5$ and keeping it at 20 otherwise. The dashed line gives $c(x)$ for both utility functions, showing

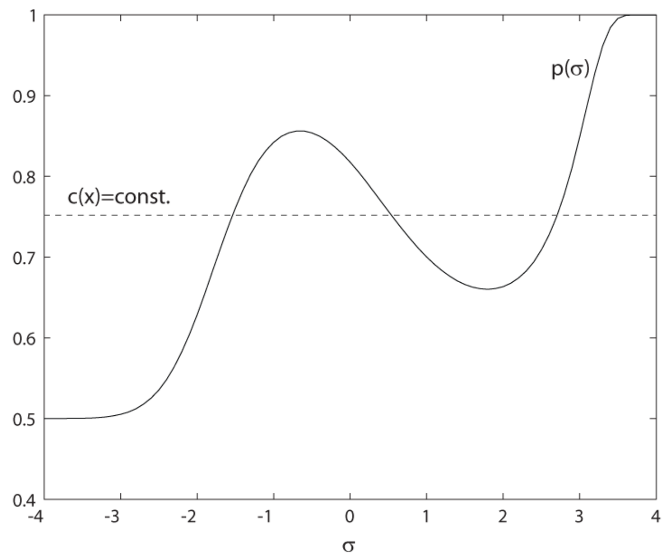

Figure 3: The solid line shows $Q(\sigma)$, the dashed line is $c(x)$ for both P3 and P4.

again three crossing points. The estimation sample has 500 observations and the out of sample evaluation 5000 observations.

Results are given in Table 2. Even though the specification of $p(X)$ is complicated, there are only three crossing points in terms of the index σ for both preferences. As Proposition 1 indicates, a cubic model in X, including all cross products, is correctly specified for the MU method. This is in contrast with ML estimation where any logit model with a finite dimensional polynomial in X will be misspecified for $p(X)$. Examining the out of sample results, all of the MU models outperform all of the ML models, even the most parsimonious MU model performs better than the best ML model. This reflects the difficulties mentioned above regarding the misspecified MLE and the ability of the MU method to still choose crossing points when the model is misspecified for $p(x)$ as a whole. The cubic MU model, which is correctly specified in the sense of capturing crossing points, is the best performer of the individual models in terms of both relative utility and also in correctly classifying the observations.

For P4 we vary the utility function to make observations towards the center of the distribution more important for utility. With this reweighting, both methods do less well in

Table 2: Model 2 - Incorrectly Specified MLE

	ML			MU			
	Lin	Cubic	Mix 1	Lin	Cubic	Mix 1	Mix 2
	In/Out	In/Out	In/Out	In/Out	In/Out	In/Out	In/Out
RU (P3)	64.8/59.8	76.5/60.1	72.5/60.1	94.2/70.9	134.7/81.4	111.2/74.6	128.0/79.6
RU (P4)	41.1/30.2	61.0/34.2	54.8/33.0	94.8/57.4	158.5/71.9	122.0/63.1	147.6/68.5

Note: The first two rows denote relative utility (RU), as defined in the first paragraph of Section 4 and expressed as a percentage, for P3 and P4. "In" refers to the in sample result, "Out" to the out of sample result.

terms of relative utility, however the loss for ML is much larger than for MU (the former dropping by about one half compared to the first utility specification). Comparing the best ML model with the best MU model for P3 (out-of-sample) we see that the ML method picks up 60% of the possible utility whereas MU picks up over to 80% of this available utility. For P4 the relative difference is larger, with ML picking up 34% and MU 72% of the available utility. This demonstrates the importance of the utility function in the semiparametric estimation procedure, and the impact it can have on the parametric methods' performance.

The mixture models here examine three specifications — the linear and cubic models (each individually reported in the table) as well as a quadratic specification (not individually reported). The selection procedure consists of testing down from the cubic specification towards the linear using the two tests described above; the three mixtures are then constructed in a similar way. As the mixture model with ML did relatively poorly when one of the ML models was correctly specified, it is no surprise that the same is true when all the ML models are not correctly specified. The gap between ML Mix 1 and MU Mix1 is considerable for P3 and even larger for P4. The method suggested in this paper (MU Mix 2) claims nearly all of the utility gained by the best single (cubic) MU model, and performs better than MU Mix 1.

Finally, rather than employing the semiparametric approach of the MU method, under lack of knowledge of the model for the conditional probability we might consider nonparametric estimation. In Table 3 we compare the two methods above to results obtained from

Table 3: ML vs. MU vs. nonparametric methods

DGP	DGP 1		DGP 2	
	P1	P2	P3	P4
Method	In/Out	In/Out	In/Out	In/Out
Nonparametric	155.3/49.2	153.2/35.4	112.9/74.8	121.1/58.8
ML(cubic)	139.4/73.1	140.9/67.1	76.5/60.1	61.0/34.2
MU(cubic)	226.0/67.1	230.9/61.3	134.7/81.4	158.5/71.9

Notes: The entries are relative utility (RU), as defined in the first paragraph of Section 4 and expressed as a percentage, for each utility function. "In" refers to the in sample result, "Out" to the out of sample result

using a nonparametric (kernel) estimator for $p(X)$. Specifically, a nonparametric regression of Y on X using the Nadaraya Watson kernel with the bandwidth chosen using least squares cross validation over the estimated sample. The results are summarized for both of the DGP's, where we have also included the best performer of the ML and MU models.

In general a consistent nonparametric estimator for $p(X)$ should attain the best utility asymptotically, however in practice we expect that its lack of focus on the crossing points will mean that in real samples it will have similar problems as other non loss based estimators. Of course we also expect that in sample problems of overfitting will be important as well. When the correct parametric specification is available, we expect that a nonparametric method would suffer from not utilizing this information. This is indeed borne out in Table 3. For the first DGP, the correctly specified ML(cubic) method has the best out of sample performance in terms of RU for both P1 and P2. With as few as 75 observations for estimation, the nonparametric method (which does not utilize the knowledge implicit in the cubic specification), performs considerably worse than ML or MU, capturing less than half of the available utility out of sample. For the second DGP, MU still outperforms the nonparametric method despite the fact that there are 500 observations in the estimation sample (X is however two dimensional). As discussed in Section 2, requiring only the knowledge of the number of turning points and the use of the utility function in the estimation of the crossing points results in a very useful method in misspecified models.

5 Conclusion

When making binary predictions or classifications with exogenous predictors, this paper provides an estimation method grounded in a general family of utility functions that has the property that, at least asymptotically, the method delivers the highest possible utility given the functional form of the model. Since the estimation procedure is geared towards fitting the optimal decision rule directly, this result does not rely on knowing the correct specification of $p(X)$. This turns out to be extremely useful when models are misspecified, a situation that is likely in practice and one where parametric models can perform very poorly. The estimation problem extends the Manski maximum score model, where the extensions are to different weighting functions and also generalizing the model beyond linear index functions. These extensions are possible since the focus is on the objective function rather than the model parameters. We present asymptotic theory showing consistency and asymptotic normality of the score function. The asymptotic results can also be used to test which of two parameterizations is closer to the maximal utility attainable given the models.

The Monte Carlo results support one of the main points of the paper — estimating a decision rule based on $p(X)$ is an exercise that requires different tools than estimating $p(X)$ itself. The typical alternative approach of employing a probit or logit model can result in very poor performance in misspecified models. Employing the utility function in the classification/forecast problem is thus shown to be important for attaining the maximal utility.

6 Appendix: Proofs

Proof (Proposition 1) Let $N = \#\Sigma < \infty$. If $N = 0$, set $h(x)$ to a constant and $A = \mathbb{R}$. If $N > 0$, let $\sigma_1 < ... < \sigma_N$ be elements of Σ. Define $K_i = 2$ if σ_i is a strict local minimum of Q and $K_i = 1$ otherwise. Setting $L(\sigma)$ equal to one of the polynomials $\pm\Pi_{i=1}^{N}(\sigma - \sigma_i)^{K_i}$ produces $sign[L(\sigma)] = sign[Q(\sigma) - \tilde{c}]$ for all σ except possibly at those σ_i at which Q is discontinuous (in this case $Q(\sigma_i) - \tilde{c} > 0$ is possible whereas $L(\sigma_i) = 0$). Let $\Sigma' \subset \Sigma$ denote the set of such σ_i. Set $h(x) = L(x'\beta)$ and $A = \{x : x'\beta = \sigma_i$, for some $\sigma_i \in \Sigma'\}^c$. Since $L(\sigma)$ is a polynomial of finite order, so is $x \mapsto h(x)$, and the set A has Lebesgue measure 1. Finally, if Q is lower semicontinuous at σ_i, then $Q(\sigma_i) \leq \tilde{c}$ so that $sign[L(\sigma_i)] = sign[0] = -1 = sign[Q(\sigma_i) - \tilde{c}]\ \forall \sigma_i$. $\blacksquare$

Lemma 1 *Under Conditions 1,2 and 3, there exists constants $K > 0$ and $q > 0$ such that, for all $\delta > 0$ sufficiently small, $\sup_{\theta_0 \in \Theta_0} E\left\{\sup_{\theta \in B(\theta_0,\delta)\cap\Theta_0} |s(Y,X,\theta) - s(Y,X,\theta_0)|^2\right\} \leq K\delta^q$, where $B(\theta_0,\delta)$ is the Euclidean open ball with radius δ centered on θ_0.*

Proof Fix $\theta_0 \in \Theta_0$ and any $\delta > 0$. By Condition 1, $|b(x)[y + 1 - 2c(x)]| \leq 4M_1$. We can write

$$E \sup_{\theta \in B(\theta_0,\delta)\cap\Theta_0} |s(Y,X,\theta) - s(Y,x,\theta_0)|^2 \leq 2(4M_1)^2 E \sup_{\theta \in B(\theta_0,\delta)\cap\Theta_0} |sign[h(X,\theta)] - sign[h(X,\theta_0)]|. \tag{7}$$

By Condition 2 part (ii), there exist constants $\lambda > 0$ and $M_2 > 0$ such that, with prob. 1,

$$h(X,\theta) \in [h(X,\theta_0) - M_2\delta^\lambda, h(X,\theta_0) + M_2\delta^\lambda] \quad \forall \theta \in B(\theta_0,\delta) \cap \Theta_0.$$

Thus for $\theta \in B(\theta_0,\delta) \cap \Theta_0$, $sign[h(X,\theta)] \neq sign[h(X,\theta_0)]$ implies that the interval above must contain zero, i.e. $h(X,\theta_0)$ must lie in $[-M_2\delta^\lambda, M_2\delta^\lambda]$. More formally,

$$|sign[h(X,\theta)] - sign[h(X,\theta_0)]| \leq 2 \times 1\left\{-M_2\delta^\lambda \leq h(X,\theta_0) \leq M_2\delta^\lambda\right\} \quad \forall \theta \in B(\theta_0,\delta) \cap \Theta_0.$$

Taking the sup of the lhs over θ, taking expectations of both sides, combining with (7), and then taking the sup over θ_0 yields

$$\sup_{\theta_0 \in \Theta_0} E \sup_{\theta \in B(\theta_0,\delta)\cap\Theta_0} |s(Y,X,\theta) - s(Y,x,\theta_0)|^2 \leq 4(4M_1)^2 \sup_{\theta_0 \in \Theta_0} P[-M_2\delta^\lambda \leq h(X,\theta_0) \leq M_2\delta^\lambda]. \tag{8}$$

By Condition 3(i), the rhs of (8) is less than or equal to $64M_1^2M_3M_2^r\delta^{r\lambda}$ for some $M_3, r > 0$ and all $\delta > 0$ sufficiently small, so $K = 64M_1^2M_2^rM_3$ and $q = r\lambda$. $\blacksquare$

Lemma 2 *Under Conditions 1, 2, and 3, the function $S(\theta)$ is (a) uniformly continuous over Θ_0, and (b) has a finite number of possible values over $\Theta_1 = \Theta \backslash \Theta_0$.*

Proof Part (a): Using Lemma 1, as $\delta \downarrow 0$ then $\sup\left\{|S(\theta) - S(\theta')| : \theta, \theta' \in \Theta_0, ||\theta - \theta'|| \leq \delta\right\}$ goes to zero. Part (b): By Condition 3(ii), for each $\theta \in \Theta_1, h(X,\theta) = h_1(X^d,\theta)$, where X^d is a sub-vector of X, and X^d has finite support. For $\theta \in \Theta_1$,

$$S(\theta) = \sum_{x^d \in \operatorname{supp}(X^d)} E\left\{b(X)[Y + 1 - 2c(X)]1(X^d = x^d)\right\} sign[h_1(x^d,\theta)].$$

By the same argument employed in the text to show $S_n(\theta)$ has finite range over Θ, this decomposition shows that $S(\theta)$ has at most $2^{\#\operatorname{supp}(x^d)}$ possible values over Θ_1. $\blacksquare$

Lemma 3 *Under Conditions 1-4* $\sup_{\theta\in\Theta_0}|S_n(\theta)-S(\theta)|\rightarrow_{a.s.}0$.

Proof We apply a modified version of Andrews (1987) to the sequence of random variables $\{s(Y_i,X_i,\theta)\}$ to obtain a uniform LLN. The modification is explained in footnote 1 of Andrews (1992). By Condition 2(i), Θ_0 is (totally) bounded. By Condition 4 and Lemma 1, Assumption A3 of Andrews (1987) holds uniformly over Θ_0; see comment 3 following Theorem 1 in Andrews (1987) and footnote 1 in Andrews (1992). By Condition 4, Assumption B1 in Andrews (1987) is satisfied. As in Lemma 1, $\sup_{\theta\in\Theta_0}|s(Y_i,X_i,\theta)|\leq 4M_1$, so Assumption B2 of Andrews (1987) is satisfied. (The measurability conditions can be argued from Appendix C of Pollard 1984). Lemma 3 then follows from Corollary 1 of Andrews (1987) and footnote 1 of Andrews (1992). ■

Lemma 4 *Under Conditions 1-4,* $\sup_{\theta\in\Theta_1}|S_n(\theta)-S(\theta)|\rightarrow_{a.s.}0$.

Proof Let $w(y,x)=b(x)[y+1-2c(x)]$. For each $\theta\in\Theta_1$, we use the decomposition of $S(\theta)$ in the proof of Lemma 2 and the analogous decomposition of $S_n(\theta)$ to write

$$\begin{aligned}|S_n(\theta)-S(\theta)| &= \left|\sum_{x^d\in\sup p(X^d)}\left\{n^{-1}\sum_{i=1}^{n}w(Y_i,X_i)1\{X_i^d=x^d\}-E[w(Y,X)1\{X^d=x^d\}]\right\}sign[h_1(x^d,\theta)]\right|\\ &\leq \sum_{x^d\in\sup p(X^d)}\left|n^{-1}\sum_{i=1}^{n}w(Y_i,X_i)1\{X_i^d=x^d\}-E[w(Y,X)1\{X^d=x^d\}]\right|,\end{aligned}$$

By the SLLN for stationary ergodic sequences (e.g. Thm 3.34 of White 2000), for any fixed value of x^d the corresponding term in $\sum_{x^d}$ goes to zero wp.1 as $n\rightarrow\infty$. As this sum is finite, the sum also goes to zero. Finally, as this upper bound does not depend on θ, $\sup_{\theta\in\Theta_1}|S_n(\theta)-S(\theta)|\rightarrow_{a.s.}0$. ■

Proof (Proposition 2). Part(a): Immediate from Lemmas 3 and 4. Part (b): For any two real-valued functions f and g defined on a set D, the inequality $|\sup_{x\in D}f(x)-\sup_{x\in D}g(x)|\leq \sup_{x\in D}|f(x)-g(x)|$ holds. Apply to $S_n(\theta),S(\theta)$ and Θ. Part (c): Using part (b)

$$\left|\sup_{\theta\in\Theta}S(\theta)-S(\hat{\theta}_n)\right|\leq\left|\sup_{\theta\in\Theta}S(\theta)-S_n(\hat{\theta}_n)\right|+\left|S_n(\hat{\theta}_n)-S(\hat{\theta}_n)\right|\leq 2\sup_{\theta\in\Theta}|S_n(\theta)-S(\theta)|,$$

which goes to zero by (a). ■

Lemma 5 *Let* $\hat{\theta}_n$ *be a measurable selection from* $\arg\max .S_n(\theta)$. *Under Conditions 1-5,* $d(\hat{\theta}_n,\Theta^*)\rightarrow_{a.s.}0$, *where* $d(.,.)$ *is the Euclidean metric.*

Proof Suppose not. Then, with positive probability, there exists $\epsilon > 0$ such that $d(\hat{\theta}_{n_j}, \Theta^*) > \epsilon$ along a subsequence $\hat{\theta}_{n_j}$. By Condition 5(ii), $\sup_j S(\hat{\theta}_{n_j}) < S^*$. However, Proposition 2(c) implies $S(\hat{\theta}_{n_j}) \to S^*$ w.p.1 as $j \to \infty$, which is a contradiction. ■

Lemma 6 *Under Conditions 1-5,* $\rho(\hat{\theta}_n, \theta^*) \to_{a.s.} 0 \ \forall \theta^* \in \Theta^*$ *where* ρ *is the semimetric on* Θ *defined by* $\rho(\theta_1, \theta_2) = \left\{E\left[|s(Y, X, \theta_1) - s(Y, X, \theta_2)|^2\right]\right\}^{1/2}$.

Proof Condition 5(i) and Lemma 2 imply $\Theta^* \subset \Theta_0$ is closed. Therefore, $\exists \theta_n^* \in \Theta^*$ such that $d(\hat{\theta}_n, \Theta^*) = d(\hat{\theta}_n, \theta_n^*)$ for each n. For these θ_n^*, for any fixed $\theta^* \in \Theta^*$, $\rho(\hat{\theta}_n, \theta^*) \leq \rho(\hat{\theta}_n, \theta_n^*) + \rho(\theta_n^*, \theta^*) = \rho(\hat{\theta}_n, \theta_n^*)$ since the $\rho-$distance between any two elements of Θ^* is zero under Condition 5(iii). Let $r_n = d(\hat{\theta}_n, \theta_n^*) + 1/n$. By Lemma 5, $r_n \to_{a.s.} 0$, so using Lemma 1, one can find n large enough so that

$$\rho(\hat{\theta}_n, \theta_n^*) \leq \left\{\sup_{\theta_0 \in \Theta} E\left[\sup_{\theta \in B(\theta_n^*, r_n)} |s(Y, X, \theta) - s(Y, X, \theta_0)|^2\right]\right\}^{1/2} \leq (K r_n^q)^{1/2}$$

for some $K, q > 0$. Letting n go to infinity completes the proof. ■

Lemma 7 *Under Conditions 1-5, the function* $J_n(\theta) = n^{1/2}[S_n(\theta) - S(\theta)]$ *is stochastically equicontinuous w.r.t. the semimetric* ρ.

Proof We verify the Assumptions of Theorem 2.2. of Andrews and Pollard (1994), where $\{s(\cdot, \cdot, \theta) : \theta \in \Theta\}$ plays the role of $\mathcal{F}$ and $\{(Y_i, X_i')\}$ plays the role of $\{\xi_i\}$. From Condition 4 the mixing coefficients satisfy the summability condition in Thm 2.2 of Andrews and Pollard (1994) with $\gamma = 2$. From Lemma 1 and the discussion preceding Theorem 2.2 the bracketing numbers of the class $\{s(\cdot, \cdot, \theta) : \theta \in \Theta\}$ are $O(x^{-\beta})$ with $\beta = 2p/r\lambda$. By the discussion following the same theorem, Condition 4 implies that the integral condition on the bracketing numbers holds with $\gamma = 2$. ■

Proof (Proposition 3). Write $n^{1/2}\left[\max_{\theta \in \Theta} S_n(\theta) - S^*\right] = n^{1/2}\left[S_n(\theta^*) - S(\theta^*)\right] + n^{1/2}\left[S_n(\hat{\theta}_n) - S_n(\theta^*)\right]$, where $\hat{\theta}_n$ is a measurable selection from $\arg\max S_n(\theta)$ and θ^* is a fixed element of Θ^*. A CLT that applies to the first term of this decomposition under the stated conditions is given by Ibragimov and Linnik (1971). $V^* > 0$ under Condition 5(iv). Further, Lemmas 6 and 7 imply $J_n(\hat{\theta}_n) - J_n(\theta^*) = o_p(1)$ or, after rearrangement, $n^{1/2}[S_n(\hat{\theta}_n) - S_n(\theta^*)] + n^{1/2}[S(\theta^*) - S(\hat{\theta}_n)] = o_p(1)$. Since $\hat{\theta}_n$ is a maximizer of S_n and θ^* is a maximizer of S, both terms of the sum are nonnegative, implying that $n^{1/2}[S_n(\hat{\theta}_n) - S_n(\theta^*)]$ must itself be $o_p(1)$.Since $\hat{\theta}_n$is a maximizer of S_n and $\theta^*$is a maximizer of S, both terms of the sum are nonnegative, implying that $n^{1/2}[S_n(\hat{\theta}_n) - S_n(\theta^*)]$ and $n^{1/2}[S(\theta^*) - S(\hat{\theta}_n)]$ are both $o_p(1)$. ■

Proof (Model selection result). Write

$$\begin{pmatrix} n^{1/2}[\hat{S}_{1n} - S_1(\theta_1^*)] \\ n^{1/2}[\hat{S}_{2n}^{\lambda} - \lambda S_2(\theta_2^*)] \end{pmatrix} = \begin{pmatrix} n^{1/2}[S_{1n}(\theta_1^*) - S_1(\theta_1^*)] \\ n^{1/2}[S_{2n}^{\lambda}(\theta_2^*) - \lambda S_2(\theta_2^*)] \end{pmatrix} + \begin{pmatrix} n^{1/2}[S_{1n}(\hat{\theta}_{1n}) - S_{1n}(\theta_1^*)] \\ n^{1/2}[S_{2n}^{\lambda}(\hat{\theta}_{2n}) - S_{2n}^{\lambda}(\theta_2^*)] \end{pmatrix},$$

where

$$\begin{aligned} S_1(\theta_1) &= E\left\{b(X)[Y+1-2c(X)]sign[h_1(X,\theta_1)]\right\} \\ S_2(\theta_2) &= E\left\{b(X)[Y+1-2c(X)]sign[h_2(X,\theta_2)]\right\}, \end{aligned}$$

and $\theta_1^* \in \arg\max S_1(\theta_1)$, $\theta_2^* \in \arg\max S_2(\theta_2)$, $\hat{\theta}_{1n} \in \arg\max S_{1n}(\theta_1)$, $\hat{\theta}_{2n} \in \arg\max S_{2n}^{\lambda}(\theta_2)$.

Using the stochastic equicontinuity argument given in the proof of Prop. 3 componentwise, the second term in the vector decomposition above is $o_p(1)$. By the multivariate CLT, the first term is asymptotically jointly normal with variance-covariance matrix

$$\begin{pmatrix} V(\theta_1^*) & \lambda E\{b^2(X)[Y+1-2c(X)]^2 sign[h(X,\theta_1^*)]sign[h(X,\theta_2^*)]\} - \lambda S_1(\theta_1^*)S_2(\theta_2^*) \\ \cdot & \lambda E\{b^2(X)[Y+1-2c(X)]^2\} - \lambda^2 S_2^2(\theta_2^*) \end{pmatrix}.$$

Suppose that $S_1(\theta_1^*) = S_2(\theta_2^*)$. Under Condition 5(iii), and the fact that the two models are nested, this is actually equivalent to $sign[h_1(X,\theta_1^*)] = sign[h_2(X,\theta_2^*)]$ w.p. 1. Furthermore,

$$n^{1/2}[\lambda\hat{S}_{1n} - \hat{S}_{2n}^{\lambda}] = \lambda n^{1/2}[\hat{S}_{1n} - S_1(\theta_1^*)] - n^{1/2}[\hat{S}_{2n}^{\lambda} - \lambda S_2(\theta_2^*)].$$

It follows that

$$n^{1/2}[\lambda\hat{S}_{1n} - \hat{S}_{2n}^{\lambda}] \longrightarrow_d N(0,\sigma^2),$$

where the asymptotic variance σ^2 is given by

$$\begin{aligned} \sigma^2 &= \lambda^2 V(\theta_1^*) + \lambda E\{b^2(X)[Y+1-2c(X)]^2\} - \lambda^2 S^2(\theta_2^*) - 2\lambda^2 V(\theta_1^*) \\ &= \lambda E\{b^2(X)[Y+1-2c(X)]^2\} - \lambda^2 S^2(\theta_2^*) - \lambda^2 V(\theta_1^*) \\ &= \lambda(1-\lambda)E\{b^2(X)[Y+1-2c(X)]^2\}. \end{aligned}$$

The first equality follows from considering the variance of a linear combination of two random variables and the stronger form of the null hypothesis. The third equality uses the null again. ■

References

[1] Andrews, D.W.K. (1987), Consistency in Nonlinear Econometric Models: A Generic Uniform Law of Large Numbers, Econometrica, 55, 1465-1471.

[2] Andrews, D.W.K. (1992), Generic Uniform Convergence, Econometric Theory, 8, 241-257.

[3] Andrews, D.W.K. and D. Pollard (1994), An Introduction to Functional Central Limit Theorems for Dependent Stochastic Processes, International Statistical Review, 62, 119-132.

[4] Boyes, W, D. Hoffman and S.Low (1989), An Econometric Analysis of the Bank Credit Scoring Problem, Journal of Econometrics, 40, 3-14.

[5] Corona, A, M. Marchesi, C. Martini, and S. Ridella, (1987), Minimizing Multimodal Functions of Continuous Variables with the "Simulated Annealing" Algorithm, ACM Transactions on Mathematical Software, 13, 262-280.

[6] Goffe, W.L., G.D. Ferrier and J. Rogers (1994), Global optimization of statistical functions with simulated annealing. Journal of Econometrics, 60, 65-99.

[7] Granger, C.W.J., 1969, Prediction with a generalized cost function, Operations Research, 20, 199-207.

[8] Granger, C.W.J. and M.H. Pesaran, 2000, Economic and Statistical Measures of Forecast Accuracy. Journal of Forecasting 19, 537-560.

[9] Horowitz, J.L. (1992), A Smoothed Maximum Score Estimator for the Binary Response Model, Econometrica, 60, 505-531.

[10] Horowitz, J.L. (1998), Semiparametric Methods in Econometrics. Springer-Verlag, New York.

[11] Ibragimov, I.A. and Y.V. Linnik (1971), Independent and Stationary Dependent Sequences of Random Variables. Wolters-Noordhoff: Groningen.

[12] Kim, J. and D. Pollard (1990), Cube Root Asymptotics, The Annals of Statistics, 18, 181-219.

[13] Lieli, R.P. and M. Springborn (2010): Closing the gap between risk estimation and decision making: efficient management of trade related invasive species risk. Working paper, Department of Economics, University of Texas, Austin.

[14] Manski, C.F. (1985), Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator, Journal of Econometrics, 27, 313-333.

[15] Manski, C.F. and T.S. Thompson (1989), Estimation of Best Predictors of Binary Response, Journal of Econometrics, 40, 97-123.

[16] Pesaran, M.H. and S. Skouras, 2002, Decision-based Methods for Forecast Evaluation. In Clements, M.P. and D. F. Hendry (Eds.), A Companion to Economic Forecasting. Blackwell, Oxford.

[17] Pollard, D. (1984), Convergence of Stochastic Processes, Springer Verag: New York.

[18] Powell, J (1994), Estimation of Semiparametric Models, In Handbook of Econometrics, Volume 4, Engle, R.F. and D.McFadden eds., North Holland: Amsterdam.

[19] Rao, R.R. (1962), Relations Between Weak and Uniform Convergence of Measures with Applications. Annals of Mathematical Statistics, 33, 659-680.