

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 8, 2005

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 and discriminant analysis. 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 formed part of Robert Lieli's dissertation at the University of California, San Diego. We have benefitted from discussions with Halbert White, Patrick Fitzsimmons, Stephen Donald, Bo Honoré, Mark Machina, Augusto Nieto Barthaburu, Max Stinchcombe and Imre Tuba. All errors are our responsibility.

1 Introduction

In a great many fields of study observations must be classified into two groups on the basis of some observed characteristics. In decision sciences, quite commonly we must make 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. In biology observations are taken and then it is determined on the basis of these characteristics if the subject is of a certain species or not. In forecasting, directional forecasts for prices are often made hence making the classification that the price goes up or they do not. In corporate finance predictions of solvency or not are based on a set of firm characteristics.

The prevalence of this problem has led to a large number of different approaches with some favored in certain fields and other methods favored in other fields. In the natural sciences methods such as discriminant analysis have been used greatly. In economics it is more likely that a logit or probit model is employed to estimate the probability of some event, with the decision being made in a two step nature following the estimation of this probability. The relative costs and benefits of various decisions and errors are typically an afterthought when it comes to estimating these probability models.

The relative costs of making errors—false negatives and false positives—are rarely balanced in the way that could be used to motivate the typical two step approaches to the problem. In detecting credit card fraud, “wasting” some time and resources on calling a customer that has full control over their credit card is not nearly as costly as failing to do so when their credit card number has been stolen. Similarly, not allowing a good candidate into a program might be quite different in cost to allowing in a poor one.

It is common in the statistical literature to introduce estimation with reference to loss functions (see Lehmann 1983), however in practice nearly all estimation proceeds through approximations to maximum likelihood methods or minimizing least squared error. In the binary decision making problem the statement of preferences, which we will refer to throughout as the utility function, takes a particularly simple form as there are only four possible categories that require enumeration.

This paper seeks to understand how models should be specified and estimated in the

framework of a full description of the utility function. The purpose of the exercise is threefold. First, to provide a general flexible framework for prediction that fully takes into account the relative weights on potential losses incurred by the decision maker. We extend the types of utility functions previously examined in this literature through allowing the utility function to depend on observed data that may be useful in determining the probability that the outcome in question would have been successful given this data. Such results are necessary to allow for decision problems that correspond to realistic situations where the point of the forecasting exercise is to take some action. Second, the general formulation will show precisely the assumptions that underly methods currently used in practice (they will be special cases, appropriate in some but not all situations). There are many papers that attempt to get at the question of which method to use, often through Monte Carlo methods or alternatively by seeing which method worked best for a particular data set. The theoretical results presented here allow a clear understanding of the features of the model that either invalidate or suggest the use of each of these special case methods. Third, the results show potential problems that can arise using common 'off the shelf' methods. The theoretical understanding gained allows us to construct experiments to show the magnitude of such problems. This paper thus unifies the literature and extends it.

The method we propose is an extension of the Manski (1975, 1985) maximum score approach. We extend the method in two ways. First we show how this method fits within the general utility approach. Second, we do not restrict ourselves to linear indexes - indeed it will be seen that linear indexes typically lack the flexibility required for complicated decision making problems. We are in a sense able to extend to nonlinear functions because of our different focus to that of Manski—we do not seek to make statements about the underlying parameters of the models, which is difficult to do in these highly nonlinear models. Instead, since we are more interested in the actual utility that arises out of these estimates, we are able to establish results that show for nonlinear specifications that we can consistently arrive at the optimal utility given the data and functional form. We establish that the rate of convergence is of the order of the square root of the sample size.

It turns out that the typical two step approach of estimating a logit or probit function

and then using the estimated probabilities to classify the data is an example of a method of forecasting where the conditional distribution of the variable to forecast is constructed and then this is employed in the decision making process. This approach, known as forecast density estimation, has garnered a large amount of attention recently with many proponents arguing that this approach is more sensible than having forecasts that are fine tuned to particular loss functions. Hence we are able in the context of a precise simple special case show the inherent pitfalls of this two step approach.

It should be noted that there are different ways to view the areas in which these results are useful. In forecasting, we can consider the decision the forecast (or alternatively that our forecasting problem admits a one to one relationship between the two possible forecasts and the two possible actions). Hence one might use the methods for forecasting bankruptcy (as often attempted in the corporate finance literature) or price change directions (as in finance). The methods could be used to forecast credit card fraud and the like. An alternative use is program design. Problems such as credit granting, admission to schools, allowing a potential participant into a program etc. also fit the general setup of the problem.

The next section describes the utility setup and optimal forecasting/classification problem in general. It is in this section that the main insights as to what is important in this problem are gained. The third section examines the estimation approach we are proposing, and establishes analytic results that suggest it will have reasonable properties in practice. In section four we review the alternate approaches. By contrasting these methods with the theory of optimal forecasting developed in the second section we are able to show the pitfalls that can occur in using these approaches. Finally, some numerical work is presented to show the magnitudes of the effects.

2 The Forecasting Framework and General Results

The binary decision we are making can be written as setting the 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. Whether or not this decision is a good one depends on some binary random variable Y , unobserved at the time

the decision is made. For example, the decision to extend the loan is good if the loan is paid back; the decision to go long is good if the price of the security is higher at the end of the holding period. Hence in this situation we set $Y = 1$ for the loan being repaid or the price going up and $Y = -1$ otherwise. This random variable is not observed at the time the loan application is reviewed or the purchase is contemplated, hence the decision maker must predict or forecast this outcome based on a number of observables. These observed data for each individual or purchase 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 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. provided that expectations exist the decision maker chooses the action to solve the maximization problem

$$\max_a EU(a, Y, X|X). \tag{1}$$

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. In a recent article in the Financial Times 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.). 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. Training programs often are required to assign individuals to a limited number of spaces in the program with the twin aims of maximizing the number of successful outcomes (say employed individuals after a job retraining scheme) but also targeting particular groups (say poorer unemployed). This results in a decision making environment where the outcome $Y = 1$ becomes the as yet unmeasured “successful completion” of the program, the action is

whether or not to enroll the individual. The covariates X enter both the prediction function as they affect the chances of success in the program and they enter the utility function of the program director due to the requirement that individuals with certain characteristics are targeted. This problem also fits time series forecasting environments. Suppose that the IMF is forecasting currency crises ($Y = 1$) and wants to take action when a country is in danger of having a crisis. The IMF utility function presumably includes not just the successful forecasting of the crisis (usually represented by the forecast error) but also characteristics of the country that make it more or less likely that the crisis will cause severe problems for international financial institutions. The same additional factors are likely to be useful for forecasting the crisis event. In each of these examples the role of the X covariates is twofold—entering through the utility function and as information useful for predicting the outcome.

Since the action and outcome are both binary, we have for any X just four possibilities. These can be described as

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

We will maintain the following assumption throughout the paper

Condition 1 (*Utility function*)

- (a) $u_{1,1}(x) > u_{-1,1}(x)$ and $u_{-1,-1}(x) > u_{1,-1}(x)$ for all x in the support of X ;
- (b) $u_{k,l}$ is Borel measurable and bounded as a function of x , i.e. $|u_{k,l}(x)| < M$ for some $M > 0$, all x and $k, l \in \{-1, 1\}$.

This condition is not restrictive and gives content to the problem. Part (a) merely states 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 . From the perspective of forecasting, this is simply the analog that making no forecast error is

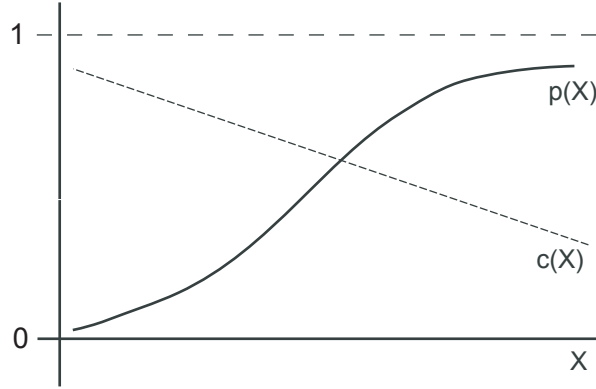


Figure 1: The basic decision/forecasting problem

better than making a forecast error, which is typically considered to be a property of loss functions (see Granger 1969). In the training program example it means that letting in those who would end up successfully completing the program is better than keeping them out, and vice versa. Moreover, this must be true for all possible realizations of the covariates X , i.e. irrespective of whether the individual is in a targeted group or not. The assumption of bounded utility in part (b) should not limit the scope of practical applications. Its main purpose is to ensure that expected values of quantities used below actually exist and to simplify some technical arguments. For most of the results presented in the paper, a second or fourth moment condition on $u_{k,l}(X)$ would actually suffice.

Denote the conditional distribution that $Y = 1$ given X as $p(X)$. In general, when $p(X)$ is known, the optimal decision can be simply calculated from this optimization problem, by integrating out the unknown value for Y . The simplicity of the structure of this particular problem leads it to admit a simple theoretical optimum. The optimal action becomes one of choosing action $a = 1$ if the conditional probability exceeds a 'cutoff' that depends on the utility function, i.e. choose $a = 1$ if and only if

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

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

gain in getting the decision correct when $Y = -1$ compared to when $Y = 1$. Hence we will choose a higher cutoff, more often taking the decision that bears fruit when $Y = -1$, as the gain from being correct in this case is larger. Thus a higher value to being correct when $Y = -1$ biases us towards forecasting this outcome more often. By construction $c(x)$ is between zero and one for any x .

Knowing the conditional probability is sufficient for making the optimal decision, one needs merely to use the utility function to decide the cutoff point. 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 (2001). The decision/forecasting problem can be pictured according to Figure 1 when X is univariate.

It is apparent from both the graph and the nature of the forecasting rule given above that the model that underlies the optimal forecast is not unique. It can be shown that the optimal problem (1) can equivalently be rewritten as

$$\max_{a(\cdot)} E_{Y,X} U [a(X), Y, X], \quad (2)$$

where the maximization is undertaken over the space of binary decision/prediction rules which are based on the observables X or, more formally, over the space of measurable functions with range $\{-1, 1\}$ defined on \mathbb{R}^k . The resulting optimal action or forecast $Y_f^* = a^*(X)$ ¹ partitions the support of X into two parts, that which corresponds to a positive action and that for a negative action. Let G^* to be the set of all (measurable) functions on \mathbb{R}^k whose *sign* produces the same partition. (We define $sign(z) = 1$ for $z > 0$ and $sign(z) = -1$ for $z \leq 0$.) Hence, we can write $Y_f^* = sign[g^*(X)]$ for $g^* \in G^*$. Similarly, every other candidate decision/prediction rule $Y_f = a(X)$ can be represented as the sign of

¹Notice that we do not distinguish between the action $a(X)$ and the point forecast $Y_f = Y_f(X)$. It is clear from Condition 1 part (a) that if the decision maker knew Y with certainty, they would always take a specific action in response, namely $a = 1$ if and only if $Y = 1$. As Y is not known, the expected utility maximizing action $a^*(X)$ is to be taken instead. We may then simply *define* the optimal point forecast Y_f^* by setting $Y_f^*(x) = a^*(x)$ for any given value x of the observed covariates X . The idea is that if the decision maker were to act trusting the optimal point forecast, he or she would be lead precisely to the expected utility maximizing action (given any value of the observed covariates). Thus, solving for the optimal point forecast is the same problem as solving for the optimal action.

suitably chosen measurable functions defined on \mathbb{R}^k . Therefore, the set of all predictors of Y can also be written in the form

$$\mathcal{P}(G) = \{sign[g(X)] : g \in G\},$$

where G is the set of measurable functions from \mathbb{R}^k to \mathbb{R} .

Rewriting (2) further and making use of the sign representation of predictors, we have that the optimization that must be undertaken is

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

where $b(X) = (u_{1,1}(X) - u_{-1,1}(X)) + (u_{-1,-1}(X) - u_{1,-1}(X))$. The role of $b(X)$ is to give higher weight to regions of X where the gain in utility is greater than other areas whether or not the outcome is positive or negative (it is the sum of the gains in each case). The optimal prediction involves choosing a candidate for $g^*(\cdot)$ amongst all possible functions, i.e. identifying the subset G^* of G . A number of points follow directly.

First, consider the situation when the the density of Y given X is known. Then from the workings above we have that one possible solution to the optimization problem is to choose $g^*(X) = p(X) - c(X)$. This solution to the problem is an element of G^* , the set of optimal predictors. It is this insight that motivates the use of a two step approach to the decision problem in practice—first estimate $p(X)$ and then examine if it is above or below the chosen cutoff point. The result shows how the cutoff point should be chosen. An examination of applications of this method shows that often the choice of the cutoff function is unrelated to any specifics of the loss function, often because the researcher has not bothered to specify the loss function in the first place.

The second point is that although full knowledge of $p(X)$ is sufficient to determine the optimal forecast Y_f^* it is not necessary. This can be most easily seen from Figure 2. Figure 2 shows the conditional probability that $Y = 1$ along with the cutoff $c(X)$ as a function of a single covariate X . As noted the optimal forecast simply involves knowing which side $p(X)$ is of $c(X)$. But consider the function $\tilde{g}(x) = m(x) - c(x)$. This function 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 that

result from using $\tilde{g}(x)$ are identical to those constructed using $p(x) - c(x)$, or more simply, $\text{sign}[\tilde{g}(x)] = \text{sign}[p(x) - c(x)]$. Hence a fundamental result in considering a model for the optimal forecast is that the models that lead to the optimal forecast are not unique. Rather than considering the existence of an optimal model it is more correct to consider that there is a set of optimal models, denoted G^* , of which $\tilde{g}(x)$ and $p(x) - c(x)$ are members. The decision maker is indifferent over the possible models since the decision remains the same.

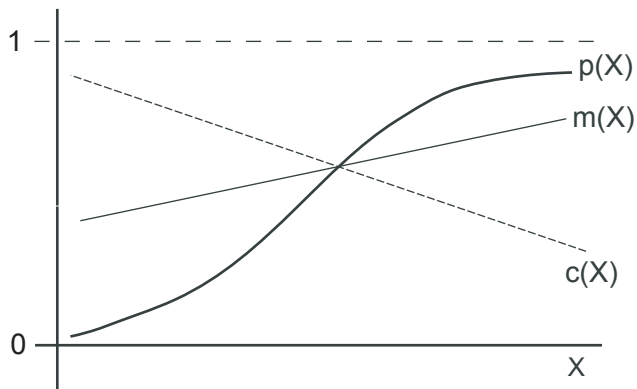


Figure 2: A possible solution to the decision/forecasting problem

The existence of a multiplicity of possible models suggests that as far as providing optimal decisions there is a greater degree of flexibility in modelling than if one were estimating $p(X)$. This third insight suggests that models that may not be all that useful as models of the conditional probability over the entire support of X —models such as the linear probability model—may still provide very good decisions by getting the “crossing point” correct. That is they may be very good approximations of the conditional probability in the regions where $p(x)$ and $c(x)$ are similar. This extra flexibility may be useful in selecting a modelling approach. This insight may also explain why in practice many researchers have found that a large number of different approaches tend to give very similar answers. This understanding also will allow us to understand when the procedures may differ.

A fourth point is that in the general problem both the optimal model and the cutoff point depends on the covariates in a potentially nonlinear manner. This means that models which are not sufficiently flexible as functions of the covariates so as to capture the possibility of multiple crossings between the cutoff and conditional probability will perform poorly in

cases where we have multiple crossings. This can be seen in Figure 3. Here $p(x)$, given by the heavier line, varies over x . But the cutoff function, given by the thinner downward sloping line, cuts this function at a number of points in x , creating three times that the function must vary in a way that leads to a change in the sign forecast. The function used must be flexible enough to capture this. Another way to see this is to consider the logit and probit when based on linear indexes (so the models are written as $\gamma(X'\beta)$). In such cases they have partial derivatives with respect to any covariate that are monotonically increasing (if the corresponding β is positive) or monotonically decreasing. Thus they have functional forms that are quite restrictive in their ability to cross the cutoff, and hence they will almost never be able to approximate an optimal model that is not monotonic. A similar story is true for maximum score estimators based on linear indexes. This suggests avoiding these methods for problems where there is likely to be more than one crossing. More flexible functional forms can avoid this problem.

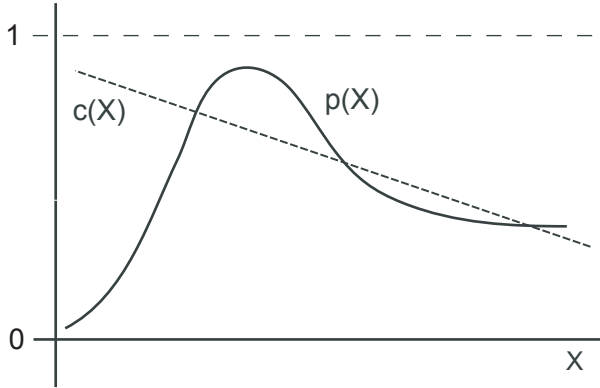


Figure 3: A non-monotonic problem

Fifth, since the “distance” between $p(x)$ and $c(x)$ is immaterial to the quality of the decision, it would seem reasonable to ignore this dimension in modelling the decision making process. Consider the (completely general) class of functions $g(x) = m(x) - c(x)$. As we have noted, so long as $m(x)$ and $p(x)$ have the properties that $sign[m(x) - c(x)]$ and $sign[p(x) - c(x)]$ are identical then the forecasts are equivalent and optimal. The distance between $m(x)$ and $c(x)$ can be removed from the optimization by simply setting $m(x)$ so that $m(x) - c(x) =$

1 when $m(x) > c(x)$ and negative one otherwise. We show this function along with $p(x)$ and $c(x)$ in Figure 4. This subset of the possible functions will include the optimal function as a point in the set of all functions of this form, hence reducing the problem in the sense of reducing the number of functions to search over and also ensuring that a single optimum exists.

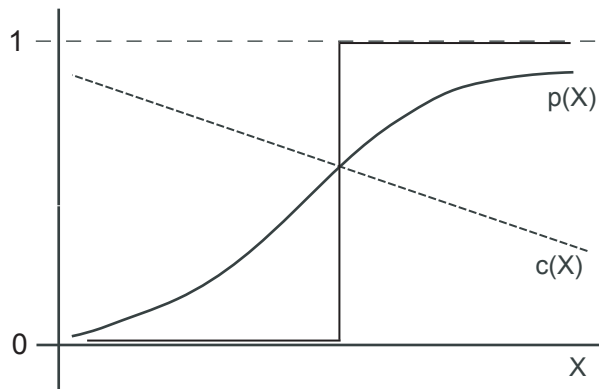


Figure 4: The step function gives $\text{sign}[m(x) - c(x)]$

A final consideration of this paper is estimation of forecasting models designed to utilize the information in the covariates efficiently give the utility function. The formulation above restricts the type of loss function that is appropriate for both estimation and evaluation. Many researchers have made the reasonable conjecture that estimation should be based on the loss function (see Weiss (1986) in the context of forecasting). A number of different loss functions have been suggested or examined. Manski and Thompson (1989) suggest the maximum score approach but rather than basing the loss function on the maximal utility consider arbitrary loss functions based on absolute and squared forecast errors. The different approaches examined in Section 4 all use loss functions that are equivalent to the family of utility based loss functions only for very special cases. For example Wang and Witten (2002) argue for minimizing Kullback Liebler distance between the true conditional probability model and the estimated one, which does not correspond in any obvious way to the loss functions that arise from utility maximization. Even more peculiar than not using the loss function suggested by the problem at hand is that many researchers estimate models

based on arbitrary loss functions and then evaluate the methods based on special cases of the loss function presented above. We will suggest sample analog methods based on the loss function, which in effect is an extension of the ideas and methods of Manski (1975, 1985) to our more general forecasting problem.

We now turn to the practical concern of estimating the model. This is undertaken in the next section.

3 Estimating Binary Prediction Models

3.1 The Maximum Utility Estimator

As we have noted the optimal forecast/allocation method chooses a function $g^*(\cdot)$ that solves

$$\max_{g \in G} E_{Y,X} \{b(X)[Y + 1 - 2c(X)]\text{sign}[g(X)]\}.$$

To actually find a solution for this optimization problem, the forecaster must search over a function space. More reasonably in practical situations the set of possible functions will be restricted in some way and a constrained optimization will be undertaken. Suppose that instead of considering all possible functions we restrict ourselves to a subset H of G and work with predictors with the form

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

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 sometimes write H_Θ for the parametrized 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

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

where the definition of $S(\theta)$ is apparent. 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 , meaning that the set of optimal

functions in the constrained problem, H^* , is also mutually exclusive of G^* . Nevertheless, if there exists $\theta^* \in \Theta$ such that $\text{sign}[h(x, \theta^*)] = \text{sign}[p(x) - c(x)]$ for all x in the support of X , then $G^* \cap H^*$ is nonempty. In standard econometric language this means that for the model H 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 $p(x) - c(x)$. Obviously, this is a much weaker requirement than fully correct specification.

To estimate a member of H_{Θ}^* , we suggest using the sample analog of the utility function that results under the model H_{Θ} , i.e. choosing θ to solve

$$\max_{\theta} S_n(\theta) = \max_{\theta} n^{-1} \sum_{i=1}^n b(x_i) [y_i + 1 - 2c(x_i)] \text{sign}[h(x_i, \theta)]. \quad (3)$$

Notice that the “reduction” in the space of functions to search over to ones that focus solely on the direction of $h(x, \theta)$ and not on the distance between $p(x)$ and $c(x)$ away from the cutoff points is built into the estimation procedure through the *sign* function. Hence the search procedure abstracts from the actual shape of say the conditional probability function and focuses the search to getting the important parts of the estimation correct, namely being on the correct side of the cutoff at the correct points. The term in front of the *sign* function, i.e. $b(x_i) [y_i + 1 - 2c(x_i)]$, is a weight assigned to each of the sign functions that is independent of the parameters of the model and differs as the utility function differs. Recall that $b(x)$ is larger when the gain from being correct is larger. Hence the weight is larger for observations which we are more interested in classifying correctly than for those which are less important to “get right”. The role of $c(x)$ is to direct the weights towards more highly valuing positive outcomes (if $c(x) < 0.5$) or negative ones (if $c(x) > 0.5$). The utility function plays a direct role in the estimation of the parameters through reweighting the observations.

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 $h(x, \theta)$ and the dimension of the parameter vector. For each θ , $S_n(\theta)$ is a sum of n terms; the absolute values of these terms do not depend on θ , only their signs do. Each component of the vector of signs

$$(\text{sign}[h(x_1, \theta)], \text{sign}[h(x_2, \theta)], \dots, \text{sign}[h(x_n, \theta)]) \quad (4)$$

is either -1 or 1 ; thus, if each component could be set independently of the others, the

whole vector would have 2^n distinct settings. However, given the shape of $h(\cdot, \cdot)$ and the realized sample points, there may not exist a value of θ to support some of these settings. Furthermore, if the form of the weights $b(x_i)[y_i + 1 - 2c(x_i)]$ is such that they are equal over many observations, there is a greater possibility that in the sum of the weighted ones and negative ones we obtain the same sum even for different settings of the sign vector (4). Thus, in practice the function S_n could take on fewer than 2^n values.

Because the range of S_n is finite over Θ , a maximum must always exist and in fact multiple maxima will exist under general conditions. Suppose θ^* solves (3) and $h(x_i, \theta^*) > 0$ or $h(x_i, \theta^*) < 0$ for each i . If $h(x, \theta)$ is continuous at θ^* , then for all θ^\dagger sufficiently close to θ^* , $\text{sign}[h(x_i, \theta^*)] = \text{sign}[h(x_i, \theta^\dagger)]$ for each i . Thus, θ^\dagger will also solve (3). This argument shows in general that S_n has “plateaus” around those continuity points of the function $\theta \mapsto h(x, \theta)$ which satisfy $h(x_i, \theta) \neq 0$ for each i . Hence, S_n is basically a step function of θ . Multiple maxima can however arise not only because the local neighborhood of θ^* is flat— S_n may very well be constant over large or disconnected regions (two different values of θ may give rise to the same sign vector even if they are not “close” and even different sign vectors could produce steps of the same height).

It is clear that maximization of S_n in practice cannot be undertaken by methods that use the gradient vector to calculate the direction of fastest ascent. Given the specification of $h(x, \theta)$, it may be possible to come up with “tricks” that facilitate the search for maxima; see Manski (1985) for the case when $h(x, \theta) = x'\theta$. In general, the simulated annealing algorithm has been shown to successfully maximize multimodal functions with flat regions or other “unpleasant” properties (Corana et al. 1987, Goffe et al. 1984). The Monte Carlo studies presented in Section 5 demonstrate that the algorithm is robust enough to handle the nonstandard nature of the objective function under consideration.

3.2 Asymptotic Properties of the Maximum Utility Estimator

Taking the utility maximization problem literally means that the primary motivation for the methods is to maximize utility and as such we are indifferent between two parameter vectors θ' and θ'' where $S(\theta') = S(\theta'')$. Multiple maxima of S are then of no concern here

as since utility is identical for all maxima the decision maker will be indifferent between the possible solutions. In such a case it makes sense that we focus more on the properties of the optimand function itself than the more usual econometrics focus on the parameters, which could be considered as nuisance parameters of the problem. We can consider two types of multiple maxima. The first occurs when the function $h(X, \theta)$ is homogenous in θ , i.e. if we can write $h(X, a\theta') = h_1(a)h_2(X, \theta')$ where $h_1(a) > 0$. In the case of a linear model this is the familiar result that θ can only be estimated up to scale, i.e. $h(X, \theta') = X\theta' = aX\theta''$ where $\theta'' = \theta'/a$ for $a > 0$ (see Manski (1985)). The second type of lack of identification occurs when the support of the covariates X is not “rich” enough to distinguish between different values for θ . This will generally occur for discrete covariates.

Each of the maxima correspond to a different estimate for the model. In many prediction situations the forecaster may be called on to justify (tell a story) about the generation of the prediction. Hence the claim above that we are not interested in the values for θ may not strictly be true. Some models may be easier to justify than others, even if all lead to the same result from a utility perspective. The same is true for program design, decisions based on one potential estimated vector may be much easier to defend than another. But this is really only a problem for the second of these types of lack of identification. In the first, although we cannot identify the actual parameters, we have that the relative marginal effects of different covariates is equivalent for all values for a , hence the “story” behind the estimation will be similar. In the second we may be able to constrain the parameter vectors in such a way as to allow for reasonable stories to be told.

A basic existence and convergence result will be proven under the following set of assumptions.

Condition 2 (*Estimation: existence and convergence*)

- (a) Θ is a compact subset of \mathbb{R}^p .
- (b) $(x, \theta) \mapsto h(x, \theta)$ is measurable with respect to $\mathcal{B}(\mathbb{R}^k) \otimes \mathcal{B}(\mathbb{R}^p)$.
- (c) (i) The function $\theta \mapsto h(x, \theta)$ is continuous on Θ for all x in the support of X .
(ii) $P[h(X, \theta) = 0] = 0$ for each $\theta \in \Theta$.

(d) $\{(Y_i, X'_i)\}_{i=1}^\infty$ is a (strictly) stationary, ergodic sequence of observations on (Y, X') .

The assumptions are quite unrestrictive. Condition 2(d) allows estimation based on dependent observations recorded over time provided that the distribution of (Y, X') remains stable. Of course the condition includes iid observations as a special case. Condition 2(c) ensures that the discontinuity of the sign function does not cause undue problems. The main restriction here is that part (ii) effectively rules out the possibility that the covariates X are all discrete and may also constrain the parameter space. For example, in the context of the linear model $h(x_i, \theta) = x'_i\theta$, it is not enough to assume that x_i has one or more continuous components. While maintaining compactness, one must also eliminate those points from the parameter space Θ that would put zero coefficients on all the continuous components (cf. Manski 1985 Assumption 2c).

First, we establish that an optimal predictor for the constrained problem exists.

Proposition 1 *Suppose that Condition 1 and Condition 2 parts (a), (b) and (c) are satisfied. Then the set $\Theta^* \equiv \arg \max_{\theta \in \Theta} S(\theta)$ is nonempty and so for any $\theta^* \in \Theta^*$ the predictor $\text{sign}[h(X, \theta^*)]$ is optimal in the class $\mathcal{P}(H_\Theta)$.*

This result establishes that there is at least one value for θ that maximizes the function S and hence can be used to construct an optimal predictor where the sense of optimality is over the constrained set of functions considered. Effectively Condition 2(c) delivers the continuity of the objective function (see Lemma 1 in the Appendix), which, coupled with the compactness of the parameter space, guarantees a maximum. Notice that whilst a maximum exists it need not correspond to a unique point θ^* in the parameter space. This is not a major concern if maximizing the utility function is the true objective of the analysis. What it will mean of course is that when there are multiple maxima of the expected utility then we will not be able to say which maximum any estimator converges to.

Now that the maximal utility $S(\theta^*)$ is well-defined, it is possible to show that the estimation procedure delivers this maximal utility asymptotically. Let $\{\hat{\theta}_n\}$ denote a sequence of estimators obtained by solving (3) for any given sample size n , i.e. $\hat{\theta}_n \in \arg \max_{\Theta} S_n(\theta)$ for each n . As the sample size tends to infinity, the estimator has the following property:

Proposition 2 *Suppose that in addition to the assumptions of Proposition 1, Condition 2 part (d) is satisfied. Then $S_n(\hat{\theta}_n) \rightarrow_{a.s.} S(\theta^*)$ and $S(\hat{\theta}_n) \rightarrow_{a.s.} S(\theta^*)$.*

The result shows that under fairly unrestrictive assumptions the classification achieved by using $\hat{\theta}_n$ converges almost surely to an optimal classification where optimality is in the sense of providing the highest possible utility given the specification of H_Θ . Such a consistency result provides a basic justification of the proposed method—it will be difficult to establish the same property for other estimators even though these other estimators are popularly used in practice.

The result relies on the function we are maximizing and the target function getting close asymptotically (see Lemma 2 and the proof of Proposition 2 in the Appendix); however, it does not rely on the actual maximized parameter vector getting close to any “true” set of coefficients. Nevertheless, in some cases it may be desirable to regard $\hat{\theta}_n$ as an estimator of some underlying truth. The conditions of Proposition 2 allow for the following, rather weak, consistency result:

Proposition 3 *Suppose Condition 1 and Condition 2 hold and let $\{\hat{\theta}_n\}$ be a sequence satisfying $\hat{\theta}_n \in \arg \max_{\Theta} S_n(\theta)$. Any convergent subsequence² of $\{\hat{\theta}_n\}$ converges almost surely to some point in $\Theta^* \equiv \arg \max_{\Theta} S(\theta)$ and, consequently, $d(\hat{\theta}_n, \Theta^*) \rightarrow_{a.s.} 0$, where $d(\cdot, \cdot)$ denotes Euclidian metric on \mathbb{R}^p .*

The proof of this result is given in the Appendix; see esp. Lemma 3. Proposition 3 implies that if one is willing to entertain additional conditions on the family of functions H_Θ and/or the distribution of (Y, X') to ensure that Θ^* consists of a single point θ^* , then $\hat{\theta}_n$ is a consistent estimator of the true optimal parameter vector θ^* . Of course, the additional assumptions needed for point identification can potentially be quite restrictive.

Proposition 3 notwithstanding, it is Proposition 2 that makes the proposed estimation method attractive from a forecasting or classification standpoint: The primary goal in this context is not the discovery of the parameter vector θ^* , but to ensure that the sample analog

²Since $\{\hat{\theta}_n\}$ is a sequence contained in the compact set Θ , a convergent subsequence is guaranteed to exist.

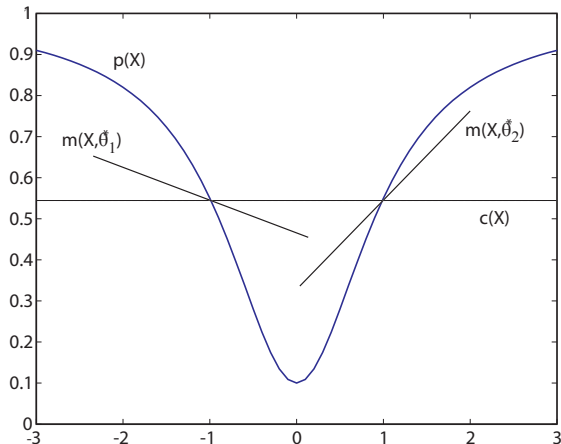


Figure 5: Maximizers of different types (X is uniformly distributed over $[-3,3]$)

of the optimal predictor attains the theoretical utility optimum in large samples under fairly weak assumptions. We will now introduce some stronger conditions under which the rate of convergence to the optimal utility level is $n^{1/2}$. More precisely, these conditions will ensure that the optimal value of the empirical objective function, $S_n(\hat{\theta}_n)$, is asymptotically normally distributed when recentered by the theoretical utility optimum and rescaled by the above rate of convergence.

In deriving the limit distribution of $S_n(\hat{\theta}_n)$, first it will be helpful to examine more closely the structure of the set $\Theta^* = \arg \max_{\Theta} S(\theta)$. Let $\theta_1^* \in \Theta^*$. Clearly, if $\theta_2^* \in \Theta$ is such that

$$\text{sign}[h(X, \theta_2^*)] = \text{sign}[h(X, \theta_1^*)] \text{ a.s.}, \quad (5)$$

then θ_2^* is also contained in Θ^* . The converse is however not true in general: if θ_1^* and θ_2^* are both in Θ^* , equation (5) does not necessarily hold. An example of this situation is depicted in Figure 5. Here X is assumed to be uniformly distributed on the interval $[-3,3]$ and the decision maker's utility function does not depend on X . The conditional probability function $p(x)$ is symmetric around zero and the class of predictors H_{Θ} is the class of affine functions in X . By the symmetry of Figure 5, we have $S(\theta_1^*) = S(\theta_2^*)$, but (5) does not hold.

Nevertheless, there is an important special case when $\theta_1^*, \theta_2^* \in \Theta^*$ implies (5). Suppose

the set of predictors H_Θ is correctly specified for the sign of $p(x) - c(x)$, i.e.

$$\text{sign}[h(X, \theta_1^*)] = \text{sign}[p(X) - c(X)] \text{ a.s. for some } \theta_1^* \in \Theta. \quad (6)$$

Then of course $\theta_1^* \in \Theta^*$ and any other maximizer θ_2^* of $S(\cdot)$ must satisfy (6) and consequently (5) as well.³

In general, for any $\theta_1^* \in \Theta^*$ one can define a set $T(\theta_1^*) \subset \Theta^*$ containing all maximizers $\theta_2^* \in \Theta^*$ such that (5) holds. Elements of $T(\theta_1^*)$ will be referred to as maximizers of the same *type* (as θ_1^*). For any two maximizers θ_1^* and θ_2^* the sets $T(\theta_1^*)$ and $T(\theta_2^*)$ must either coincide or be disjoint, so the “type-sets” form a partition of Θ^* (each maximum has a unique type).⁴

We will establish the asymptotic distribution of $S_n(\hat{\theta}_n)$ under the assumption that all maximizers of $S(\theta)$ are of the same type. In addition, we will need to strengthen the restrictions Condition 2 places on $h(X, \theta)$ and the sequence $\{(Y_i, X_i')\}$.

Condition 3 (*Rate of convergence and asymptotic distribution*)

(a') Θ is a compact subset of \mathbb{R}^p .

(b') $(x, \theta) \mapsto h(x, \theta)$ is measurable with respect to $\mathcal{B}(\mathbb{R}^k) \otimes \mathcal{B}(\mathbb{R}^p)$.

(c') (i) The function $\theta \mapsto h(x, \theta)$ is Lipschitz-continuous on Θ , i.e. there exists a function $L : \mathbb{R}^k \rightarrow (0, \infty)$ and $\lambda > 0$ such that for all $\theta, \theta' \in \Theta$

$$|h(x, \theta) - h(x, \theta')| \leq L(x) \|\theta - \theta'\|^\lambda.$$

(ii) There exists a component $X^{(j)}$ of X such that the distribution of the random variable $h(X, \theta)/L(X)$ conditional on $X^{(1)}, \dots, X^{(j-1)}, X^{(j+1)}, \dots, X^{(k)}$ is absolutely continuous w.r.t. Lebesgue measure for all $\theta \in \Theta$, and the thus existing conditional density of $X^{(j)}$ is bounded in a neighborhood of zero, where both the neighborhood and the bound are uniform in $x^{(1)}, \dots, x^{(j-1)}, x^{(j+1)}, \dots, x^{(k)}$ and θ .

³If (6) is violated by some $\theta^\circ \in \Theta$, then the predictor $\text{sign}[h(X, \theta^\circ)]$ contradicts $\text{sign}[p(X) - c(X)]$, the optimal predictor, with positive probability (i.e. on an x -set with positive F_X -measure). By Condition 1(a), this will lead to a nonzero loss in expected utility. Hence, θ° could not be a maximizer of $S(\cdot)$.

⁴Using Condition 2(c) and an argument similar to the proof of Lemma 1 in the Appendix, it is possible to show that the set $T(\theta^*)$ is closed (and hence compact) for any $\theta^* \in \Theta^*$. Therefore, if Θ^* consists of a finite number of types of maximizers, then the subsets containing the different types are disconnected.

(d') $\{(Y_i, X'_i)\}_{i=1}^\infty$ is a (strictly) stationary, strong mixing sequence of observations on (Y, X') with mixing coefficients $\alpha(d)$ that satisfy

$$\sum_{d=1}^{\infty} d^{Q-2} \alpha(d)^{\gamma/(Q+\gamma)} < \infty \quad (7)$$

for some even integer $Q \geq 2$ and some $\gamma > 0$ such that $Q/(2+\gamma) > p/\lambda$. The numbers p and λ are defined in parts (a') and (c')(i), respectively.

Parts (c') and (d') of Condition 3 are stronger than the assumptions made in the corresponding parts of Condition 2. Lipschitz-continuity clearly implies the pointwise continuity assumed in Condition 2 part (c)(i). Furthermore, Condition 3 part (c')(ii) can be regarded as a strengthening of the corresponding part of Condition 2. To see this, suppose that the function $L(x)$ is identically equal to one. Then Condition 3 requires $h(X, \theta)$ itself to have a conditional density. Assuming $X^{(j)} \equiv X^{(1)}$, we can write, for all $\theta \in \Theta$,

$$\begin{aligned} \mathbb{P}[h(X, \theta) = 0] &= E[1_{\{h(X, \theta)=0\}}] \\ &= E\{E[1_{\{h(X, \theta)=0\}} \mid X^{(2)}, \dots, X^{(k)}]\} \\ &= E\{\mathbb{P}[h(X, \theta) = 0 \mid X^{(2)}, \dots, X^{(k)}]\} = E\{0\} = 0. \end{aligned}$$

Just as its weaker counterpart, Condition 3 part (c')(ii) essentially requires the presence of a continuous covariate and also restricts the specification of $\{h(\cdot, \theta) : \theta \in \Theta\}$. For example, the function $x \mapsto h(x, \theta)/L(x)$ must be nontrivial in the continuous argument $x^{(j)}$ for any value of θ and $x^{(1)}, \dots, x^{(j-1)}, x^{(j+1)}, \dots, x^{(k)}$.

Condition 3 part (d') restricts the memory of the sequence $\{(Y_i, X'_i)\}$; the restriction is stronger for larger values of Q and smaller values of γ . Therefore, ceteris paribus, one would want to choose as small a Q and as large a γ as possible without violating the constraints $Q/(2+\gamma) > p/\lambda$ and $Q \geq 2$. As far as the tradeoff between decreasing Q and increasing γ is concerned, Q and γ can be chosen optimally in the following way. Let $\alpha(d) = O(d^{-a})$ for some $a > 0$. Then condition (7) requires $Q - 2 - a\gamma/(Q + \gamma) < -1$ or $a > Q^2/\gamma - \gamma$. To make the memory condition as weak as possible, we need to choose a as small as possible, which can be achieved by solving the constrained optimization problem

$$\min_{\gamma, Q} (Q^2/\gamma - \gamma) \text{ s.t. } Q/(2+\gamma) > p/\lambda, Q \geq 2, \gamma > 0.$$

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

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

where $s(Y, X, \theta) \equiv b(X)[Y + 1 - 2c(X)]\text{sign}[h(X, \theta)]$. It is clear that if θ_1^* and θ_2^* are maximizers of the same type, then $V(\theta_1^*) = V(\theta_2^*)$. In the special case when $\{(Y_i, X'_i)\}$ is an i.i.d. sequence, $V(\theta)$ reduces to

$$V(\theta) = \text{var}[s(Y_1, X_1, \theta)] = E\{[b(X)]^2[Y + 1 - 2c(X)]^2\} - S(\theta).$$

Hence, in this case the asymptotic variance is the same for any maximizer $\theta^* \in \Theta^*$, regardless of type.

We have the following result:

Proposition 4 *Suppose Condition 1 and Condition 3 are satisfied and all the elements of $\Theta^* = \arg \max_{\Theta} S(\theta)$ are of the same type. Let S^* and V^* denote the value of $S(\cdot)$ and $V(\cdot)$ over Θ^* . Moreover, let $\hat{\theta}_n$ be a sequence of parameters satisfying $\hat{\theta}_n \in \arg \max_{\Theta} S_n(\theta)$. If $V^* > 0$, then*

$$n^{1/2}[S_n(\hat{\theta}_n) - S^*] \longrightarrow_d N(0, V^*).$$

The proof of Proposition 4 is based on the decomposition

$$n^{1/2}[S_n(\hat{\theta}_n) - S^*] = n^{1/2}[S_n(\theta^*) - S(\theta^*)] + n^{1/2}[S_n(\hat{\theta}_n) - S_n(\theta^*)], \quad (8)$$

where θ^* is an arbitrary (but fixed) element of Θ^* . As θ^* is fixed, the first term on the RHS of (8) has a normal limit distribution. This can be established by applying a central limit theorem to the sequence of random variables

$$s(Y_i, X_i, \theta^*) = b(X_i)[Y_i + 1 - 2c(X_i)]\text{sign}[h(X_i, \theta^*)], \quad i = 1, 2, \dots, n.$$

The sequence $s(Y_i, X_i, \theta^*)$ inherits stationarity and mixing from (Y_i, X'_i) , and the mixing coefficients of the transformed sequence decay at least as fast as those of the original. As $s(Y_i, X_i, \theta^*)$ is a sequence of bounded random variables, fairly weak memory conditions suffice

for a CLT (see, e.g., Ibragimov and Linnik 1971, Thm. 18.5.4). Namely, it is enough for the mixing coefficients of (Y_i, X_i') to be summable, which is more than guaranteed by Condition 3 part (d). Thus, the aforementioned CLT can be applied to the object

$$n^{1/2}[S_n(\theta^*) - S(\theta^*)] = n^{-1/2} \sum_{i=1}^n \{s(Y_i, X_i, \theta^*) - E[s(Y, X, \theta^*)]\}$$

to establish a normal limit distribution with variance V^* .

The second term on the RHS of (8) is $o_p(1)$ under Condition 3 and hence has no bearing on the limit distribution. This is a consequence of the stochastic equicontinuity of the function

$$J_n(\theta) \equiv n^{1/2}[S_n(\theta) - S(\theta)]$$

with respect to the semimetric

$$\begin{aligned} \rho(\theta_1, \theta_2) &= \left\{ E \left[|s(Y, X, \theta_1) - s(Y, X, \theta_2)|^2 \right] \right\}^{1/2} \\ &= \left\{ E \left[|b(X)Y + 1 - 2c(X)|^2 |\text{sign}[h(X, \theta_1)] - \text{sign}[h(X, \theta_2)]|^2 \right] \right\}^{1/2}. \end{aligned}$$

The function $\rho(\cdot, \cdot)$ has all the properties of a metric except that $\rho(\theta_1, \theta_2) = 0$ does not imply $\theta_1 = \theta_2$. In particular, if θ_1^* and θ_2^* are two maximizers of $S(\cdot)$ having the same type, then $\rho(\theta_1^*, \theta_2^*) = 0$ even when $\theta_1^* \neq \theta_2^*$. Loosely speaking, the “same-type” assumption in Proposition 4 makes the set of maximizers Θ^* behave like a single point under the semimetric ρ , eliminating the difficulties stemming from the existence of multiple maximizers.

Stochastic ρ -equicontinuity of $J_n(\cdot)$ follows from Theorem 2.2 of Andrews and Pollard (1994) (see the Appendix for details) and is employed in the following way. By Proposition 4, $d(\hat{\theta}_n, \Theta^*) \rightarrow_{a.s.} 0$, where $d(\cdot, \cdot)$ denotes Euclidian distance. By Lemma 4 in the Appendix, this implies $\rho(\hat{\theta}_n, \theta^*) \rightarrow_{a.s.} 0$ for *any* fixed $\theta^* \in \Theta^*$. Stochastic equicontinuity w.r.t. ρ then gives $J_n(\hat{\theta}_n) - J_n(\theta^*) = o_p(1)$ or, after rearrangement,

$$n^{1/2}[S_n(\hat{\theta}_n) - S_n(\theta^*)] + n^{1/2}[S(\theta^*) - S(\hat{\theta}_n)] = o_p(1). \quad (9)$$

Since $\hat{\theta}_n$ is a maximizer of S_n and θ^* is a maximizer of S , both terms of the sum on the LHS of (9) are nonnegative, implying that $n^{1/2}[S_n(\hat{\theta}_n) - S_n(\theta^*)]$ must itself be $o_p(1)$ as claimed.

The fact that Proposition 4 relies on a fairly “high-level” assumption (namely that all maximizers of $S(\cdot)$ are of the same type) is somewhat of a shortcoming. The discussion

preceding Proposition 4 shows that this assumption is guaranteed to hold when $h(x, \theta)$ is parameterized richly enough so that condition (6) is satisfied. It is however difficult to see what precise restrictions on $h(x, \theta)$ and the distribution of (Y, X') are required to make the “single type” assumption generally true when $h(x, \theta)$ is underparameterized. The example presented in Figure 5 suggests that rather special circumstances would have to be in place for maximizers of different types to exist, at least when X is a scalar. While it is not clear if this is still true when X is higher dimensional, the “single type” assumption is, at minimum, a reasonable starting point.⁵

3.3 Model Selection Based on Maximum Utility Estimation

As an application of Proposition 4, we outline the development of a model selection criterion useful for finding the best specification for $H_\Theta = \{h(\cdot, \theta) : \theta \in \Theta\}$ among various candidates. The construction will only depend on the fact that $n^{1/2}[S_n(\hat{\theta}_n) - S^*]$ is $O_p(1)$ but not on the exact form of the limit distribution.

In particular, let $\{h(\cdot, \theta_i) : \theta_i \in \Theta_i\}$, $i = 1, 2, 3$, represent three nested model specifications.⁶ The dimensions of the parameter vectors are p_1, p_2 and p_3 , respectively. We will refer to the first model as underfit, the second as the (pseudo) true model and the third as an

⁵Even if $S(\cdot)$ has maximizers of different types, Proposition 4 may remain entirely valid or require only minor modifications. Suppose there are two types of maximizers belonging to the sets Θ_1^* and Θ_2^* , respectively. By Proposition 3, almost all realizations of the sequence $\{(Y_i, X'_i)\}$ are such that $d(\hat{\theta}_n, \Theta^*) \rightarrow 0$ whenever $\hat{\theta}_n \in \arg \max S_n(\theta)$ for each n . It is then possible to show that $\hat{\theta}_n$ can be decomposed into two disjoint (and collectively exhaustive) subsequences $\hat{\theta}_{n_1(s)}$ and $\hat{\theta}_{n_2(s)}$ such that $d(\hat{\theta}_{n_1(s)}, \Theta_1^*) \rightarrow 0$ and $d(\hat{\theta}_{n_2(s)}, \Theta_2^*) \rightarrow 0$ as $s \rightarrow \infty$. (One of the subsequences may be “empty”.) The indices $n_1(s)$ and $n_2(s)$ are random: they depend, in general, on the realization of the sequence $\{(Y_i, X'_i)\}$ (and also on how the $\hat{\theta}_n$ are selected from $\arg \max S_n(\theta)$); that is, one should write $n_1(s) = n_1(s, \omega)$ and $n_2(s) = n_2(s, \omega)$. Still, it should be fairly straightforward to justify the use Proposition 4 to evaluate the asymptotic distribution of $n^{1/2}[S_n(\hat{\theta}_n) - S^*]$ along the subsequences $n_1(s)$ and $n_2(s)$. In the i.i.d. case the resulting limit distribution is $N(0, V^*)$ along both subsequences, where V^* is the common value of $V(\theta)$ over Θ_1^* and Θ_2^* . Hence, the limit distribution of $n^{1/2}[S_n(\hat{\theta}_n) - S^*]$ itself is given by the same law, being some mixture of two identical distributions. It should be possible to make this argument fully rigorous and extend it to the stationary mixing case as well.

⁶In this section we will abuse notation and write $h(x, \theta_1)$, $h(x, \theta_2)$, etc instead of $h_1(x, \theta_1)$, $h_2(x, \theta_2)$, etc.

overfit model.

The *pseudo true* parametrization is characterized by the property that there exists $\theta_2^* \in \Theta_2$ such that $\text{sign}[h(X, \theta_2^*)] = \text{sign}[p(X) - c(X)]$ a.s., so $S(\theta_2^*)$ attains the unconstrained utility maximum. Furthermore,

- an *underfit* is defined by $p_1 < p_2$ and $S(\theta_1^*) < S(\theta_2^*)$ for $\theta_1^* \in \arg \max S(\theta_1)$, so the unconstrained utility maximum is not achieved;⁷
- an *overfit* is defined by $p_3 > p_2$ and $S(\theta_2^*) = S(\theta_3^*)$ for $\theta_3^* \in \arg \max S(\theta_3)$, so the unconstrained utility maximum is attainable by both models, but the overfit model has a greater number of parameters.

Consider a model selection criterion of the form

$$M_n(\hat{\theta}_{i,n}) = S_n(\hat{\theta}_{i,n}) - p_i g(n),$$

where $\hat{\theta}_{i,n} \in \arg \max_{\theta_i \in \Theta_i} \hat{S}_n(\theta_i)$. We wish to specify the penalty term $p_i g(n)$ so that the criterion achieves, asymptotically, a strict maximum under the pseudo true parametrization. The penalty term is necessary because by the properties of maximization $\hat{S}_n(\hat{\theta}_{i,n})$ will always be the largest under the most general specification.

The following result is based solely on the fact that $n^{1/2}[S_n(\hat{\theta}_{i,n}) - S(\theta_i^*)]$ is $O_p(1)$. The proof can be found in the Appendix.

Proposition 5 *Suppose Condition 3 is satisfied. Let $g(n) = C \log(n)/\sqrt{n}$ for some positive constant C . Then*

$$P[M_n(\hat{\theta}_{2,n}) > M_n(\hat{\theta}_{i,n})] \rightarrow 1$$

as $n \rightarrow \infty$ for $i = 1$ (*underfit*) and $i = 3$ (*overfit*).

A number of comments are in order.

1. It is apparent from the proof of Proposition 5 that any choice of $g(n)$ such that $g(n) \rightarrow 0$ and $n^{1/2}g(n) \rightarrow \infty$ would yield a consistent selection criterion.

⁷One could imagine underfitting with more parameters than the optimal model, but here we are concentrating on nested specifications.

2. The form of the penalty term does not have the usual foundation here; the use of p_i is rather *ad hoc*. Typically the penalty arises as the mean of a random variable describing the addition to the sampling error by overfitting. We have not derived this random variable.

3. The consistency of the model selection criterion M_n is by any means a desirable property, but it still leaves open the practically important question of how “best” to choose the constant C (more generally, the penalty for overfitting) in finite samples. If C is chosen to be “too small” for a given sample size n , then richer parameterizations will be accepted “too often”, even when the increase in the maximized value of the empirical score is only due to finite sample variation and the model’s increased ability to adapt to these random patterns. On the other hand, if C is chosen to be “too large” for a given sample size n , then richer parameterizations will be rejected “too easily”, even when the larger model leads to some increase in the maximum of the *true* score function as well.

4 Alternative Methods for Estimating Decision Functions

The asymptotic results presented in Sections 3 provide grounds for using the utility based estimator in practice. In most applications other methods have been employed. We now examine these other methods.

One of the most popular methods is the use of logit or probit models to estimate $p(X)$, followed by the use of an arbitrary (non loss function) motivated cutoff choice. One could also of course use linear probability models to approximate the conditional probability. Other models and estimation techniques for the conditional probability such as neural nets have also been entertained. Popular in biological sciences and also corporate finance, the method of discriminant analysis first introduced by R.A. Fisher (1936) can be used for this problem. Finally, the Manski maximum score method is closely related to the method suggested above—it is a special case.

4.1 Parametric Estimation of $p(X)$

The most common approach to estimating the conditional probability is through the estimation of a probit or logit regression of Y on X . These methods estimate the conditional probability that Y takes the value 1 through choosing a parametric function $\gamma(X'\beta)$ for the model of the conditional probabilities where β are unknown parameters to be estimated from the available dataset. Thus under the assumption that the cutoff is chosen appropriately the set of models considered here can be written in the form

$$\{\text{sign}[\gamma(X'\beta) - c(X)], \beta \in \mathbb{R}^k\}$$

Typically the logit specification for $\gamma(\cdot)$ is preferred over probit in application. Dimitris et. al (1996) survey business failure prediction models and find that 17 of 66 of the studies surveyed employed this method. Boyes et. al. (1989) use these logit models to predict credit default. Leung et. al (2000) use both probit and logit models to predict the direction of the stock market. Martin (1977) and Ohlsen (1980) use this approach to predict the bankruptcy of corporations. The choice of a cutoff probability — above which we assign a forecast of a one — tends to be arbitrary in these studies. Leung et. al. (2000), Min and Sha (2003) each choose one half as the cutoff. Boyes et. al (1989) suggest a loss based cutoff.

First, suppose that the parametric form for the conditional probability is correctly specified up to the unknown parameters β . In this case the MLE $\hat{\beta}$ is consistent for β and hence $\hat{p}(x)$ converges to $p(x)$ in some suitable sense. Provided that the correct cutoff function is used then $p(x) - c(x) \in G^*$ and so asymptotically these methods are able to provide the correct decision rule. This all happens despite the fact that the implicit loss function underlying the estimation procedure does not correspond to the actual loss function for the problem except in very simple cases. This is an unusual coincidence, and follows from the fact that $p(x)$ plays two roles in this particular problem, one as the reduced form statistic required to obtain the forecasts and secondly it is the predictive density. In its second role, the loss function is not required asymptotically since we are able to estimate the density and then construct the decision. In the first role, it is one of the possible sufficient statistics for the problem. This does not occur for most decision rule problems, typically the forecast

based on an incorrect loss function is not asymptotically equivalent to knowing the predictive density.

In the more likely situation that the parametric model is misspecified, there is no reason why the MLE should asymptotically find a function in H^* even when the cutoff is correctly specified. The intuition for this follows from considering what the maximum likelihood procedure does versus what is required for a good decision. The method of maximum likelihood will attempt to find the best global fit for $p(\cdot)$. So the method will attempt to fit the true conditional probability not only at the points where it matters for decision making—namely at the cutoff point—but also at other irrelevant points where good decision making requires only the side of the cutoff and information about the conditional probability other than this is not informative. Depending on the nature of the conditional probability function and the densities of the data it may well be that the method of maximum likelihood places all its effort on fitting the conditional probability at points far from the cutoff. Figure 6 depicts with the solid line $p(x)$ for a single covariate. No probit or logit model based on a linear index can approximate this well over the entire range of x . Here, with X uniformly distributed, the population probit estimates lead to the lighter shaded dashed line. We see that it does provide a good approximation at a number of points (i.e. the points at which it crosses $p(x)$ and is hence equal to the conditional probability), however at many other points it does not. If the cutoff happened to be at 0.5, the method would have done well. If it is at 0.4 (as in Figure 6 denoted by the horizontal dotted line) then it is a very poor approximation.

Finally, parametric models for the conditional probability may not be flexible enough to capture the number of crossing points, i.e. points where $p(x)$ cuts $c(x)$. This is a particular type of misspecification, where the optimal classification scheme may break up the region over the covariates into more parts than the parametric form of the model will allow. The models may then be not flexible enough to optimally classify over all regions. This problem can be seen in the Figure 7, where now $p(x)$ first rises then falls to an asymptote. Any cutoff independent of x between about 0.38 and 0.81 will result in two crossings of c with $p(x)$. However parametric models based on a linear index are monotone, hence will not be able to

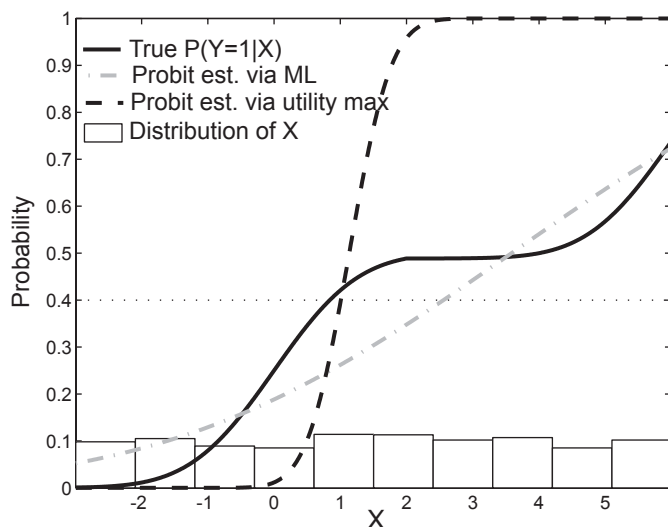


Figure 6: A misspecified probit model

pick both points. They could pick two points, however this would mean that at least one point for the crossing is misspecified.

4.2 Nonparametric Estimation of $p(X)$

An alternative approach to probit or logit estimation is to construct an estimate of $\hat{p}(x)$ using less parametric methods. Use of probit or logit results in models that are restricted in the form of the function $G(\cdot)$. The motivations for less parametric methods are driven by the lack of any basic theoretical basis for a particular model specification of $p(x)$ and a desire to avoid the problems discussed above when the parametric form is prespecified. A large number of less parametric approaches are available. Many of the available methods arise from methods designed to semi or nonparametrically estimate β (which is not the direct problem of the forecaster, but arises in many other areas). With a consistent estimate for β then the functional form can be estimated conditionally on the parameter vector estimate (Powell 1994 has a review of these methods). Such methods have not to our knowledge been employed for forecasting. Most semi or nonparametric methods actually employed for forecasting tend to be applications of neural network methods or other model search procedures. Examples in predicting bankruptcy include Wilson and Sharda (1994). Theoretically then these methods

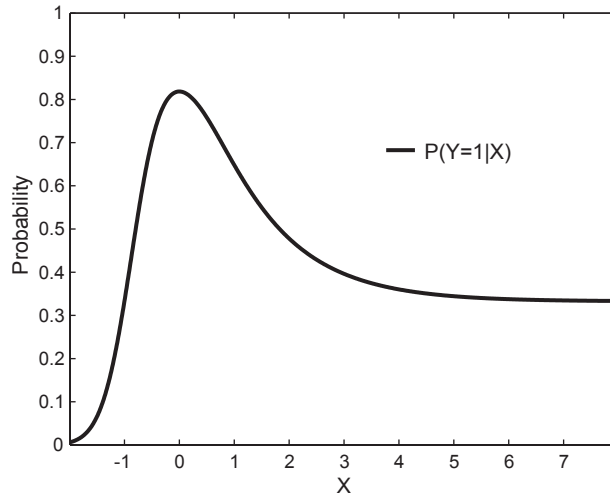


Figure 7: An example of $p(x)$ where there may be multiple “crossings”

are more likely to avoid the problem of specifying a set of possible models that excludes the optimal model, at the cost empirically of the need for greater model search and the potential for overfitting.

4.3 Linear Probability Model

The linear probability model simply fits the usual regression of \tilde{Y} on X where \tilde{Y} is equal to Y for positive outcomes and zero otherwise. It does this without regard to the binary nature of the dependent variable and the sensible requirement that the model not predict numbers other than one or zero. Because of the nature of the dependent variable the least squares residuals must be heteroskedastic, so it is typical to employ weighted least squares based on initial OLS estimates.

This method has not been too often used in the prediction problem. The linear probability model cannot provide a good fit of the entire conditional probability $p(x)$. However as we have demonstrated above, for it to provide useful predictions it need only give a good estimate of the conditional probability at the point where it cuts the cutoff function $c(x)$. Hence we may well expect it to provide good forecasts in some situations. As with the other methods described in this section above, it fails to utilize any information in the utility function and

hence whether or not the performance of this method is good or bad is difficult to determine a priori.

4.4 Discriminant Analysis

The method of discriminant analysis splits the joint distribution of the X covariates into two different populations, one group for when $Y = 1$ and the other for when $Y \neq 1$. The derivation of this method arises from considering a hypothesis test between these two covariate populations. Defining these population joint densities parametrically, we can compute the likelihoods that any subsequent observed set of covariates are generated from the density associated with the $Y = 1$ group or the alternative group. A likelihood ratio test between two groups results in the rule of forecasting that $Y = 1$ if the likelihood evaluated for this group at $X = x$ exceeds that for the other group. When the assumption that X is jointly normally distributed with common variance covariance matrices across groups is made, this rule is linear in the observed x 's and hence leads to decisions based on linear 'scores' that have a one to one correspondence with decisions based on the underlying likelihoods. Rather than weight both likelihoods evenly, the researcher can weight one higher than the other resulting in a procedure that alters the balance between false positives and false negatives made by the forecast procedure. In this way the loss function can be brought to bear on the decision making through these weights. It may seem odd that the focus of these methods is on the distribution of the covariates rather than the conditional distribution of the outcome variable to be forecast — the methods are best understood as getting to this conditional distribution via the complete joint distribution of Y and X .

This method is often criticized because of their parametric assumptions (joint normality of the covariates) however this really amounts to the same thing as choosing a parametric model for the conditional probability. This becomes clear through their relationship with the logit model. Amemiya (1985) shows that under the normality assumption for the distribution of X conditional on Y that the discriminant analysis model is equivalent to a logit model that includes the X covariates linearly but also quadratically. If the assumption that the variance covariance matrices are equivalent for the sub populations then the quadratic terms

drop out. There is in population a one to one relationship between the logit specification based on a linear index and the method of discriminant analysis when the variances are assumed equal. Discriminant analysis methods have been popular in statistics and applied fields other than economics, and date back to work by R.A. Fisher (1936). For reviews of the methods in general (as opposed to forecasting) see Amemiya (1981) or Maddala (1983). Since for the special (but common in applications) case where the variance covariances of the two covariate populations are held to be the same decisions based on discriminant analysis and decisions based on linear probability models will both be based on linear functions of the data. Dimitris et. al (1996) note the similarity of predictions using these two methods, where differences arise due to different estimation methods for the unknown parameter weights.

This method has been employed for detecting credit granting (Srinivasan, V. and Y.H. Kim (1987)) and bankruptcy prediction (Dimitras et. al. (1996)) amongst many other applications.

4.5 Manski Maximum Score Approach

The Manski maximum score method (Manski (1985), Manski and Thompson (1989)) also fits into this set of approaches. Define the 'score' as the proportion of correct matchings between the model and the binary outcomes (for any set of parameters chosen, the model gives only two outcomes and hence is a step function). The method here is to choose parameters so that the score is maximized. The estimation procedure described above is a generalization of the one introduced in Manski (1985), which is identical to the above approach when (i) $b(x)$ and $c(x)$ are independent of x and (ii) $h(x, \theta) = x'\theta$.

This method has been less used in forecasting than other methods⁸ described above, which is somewhat strange since in many applications of the above methods researchers then often compare models based on the proportion of correct predictions — i.e. they use as a performance criterion the exact loss function that the most common form of the maximum score estimator maximizes. One potential reason for its underuse is that evaluating the

⁸Indeed, we know of only one application. Caudill (2003) uses the method to predict the outcome of basketball games.

sample distribution of the estimated parameters is somewhat difficult due to the discontinuous objective function (Horowitz (1992)), and whilst this is a problem for model selection it is not a problem for estimation and actually using the model to construct forecasts is straightforward.

5 Large Sample Properties of Classification Techniques

We will examine a number of the 'population' effects discussed above using simulations with large samples. We will demonstrate how failure to utilize the utility function in misspecified models results in inferior classification when using two step procedures, showing the situations in which this is more of a problem or less of a problem. We will first start with linear models, typically the more popularly used type of specification. We will then give a nonlinear example.

When the model for the conditional probability $p(X)$ is correctly specified then in large samples classifications using the correct cutoff will of course maximize utility. As we have discussed above there is then no difference between the two step procedure and the one step procedures in this special case. Because of this, in this section we focus on a situation where the conditional probability is not correctly specified by any of the models. The data generating process is given by

$$Y = \text{sign}[1 + X + e^X v]$$

where X is a univariate scalar random variable and the innovation term v is mean zero and independent of the covariate X . In particular we model the innovation as being lognormal, i.e. $v \sim \exp[N(0, \sigma^2)] - \exp(\sigma^2/2)$. The conditional probability is given in Figure 6 above.

For this data generating process, linear probit or logit models based on the index $\beta_1 + \beta_2 x$ will be misspecified, they are not flexible enough to capture the features of the conditional probability over the full range of conditional probabilities. This does not mean that they are not able to capture some of the features of the conditional probability, maximum likelihood estimation will attempt to fit this curve as well as possible over the range of observed X variables.

Since the ability of the two step methods for approximating the conditional probability depend on the relevant range over which the approximation is made, we will consider three distributions of the X covariate to bring out the points to be made. The distributions are (i) $X \sim \chi_2^2 - 2$, (ii) $X \sim U(-2, 8)$ and (iii) $X \sim N(2, 1)$. The first of these concentrates the mass of X at the far left of Figure 6, the second spreads it evenly across the horizontal axis, and the third places most mass in the center and less at each end of the horizontal axis. From the perspective of the estimation methods that do not use information in the utility function (i.e. ignore where the cutoff point will be in the classification procedure), this corresponds to the methods attempting to approximate $p(X)$ over the initial upswing for the first distribution for X , the whole curve for the second distribution and the middle of the graph for the third distributional choice.

How useful these approximations will be (and how costly the mistakes will be) depend on the utility function. We will employ a number of different assumptions. For the initial set of demonstrations, we simplify the utility function so that it does not depend on X and hence the cutoff function will be a constant (a horizontal line in the Figure). For utility functions that lead to cutoff points between about 0.35 and 0.8, there will be three regions for the classification—the first and third being where $p(x) < c(x)$ with a middle region over x where the reverse is true. Hence a single linear index will not have any chance of getting all of the regions correct. With $c < 0.35$ there are only two regions and hence it is possible for the misspecified model to cleave the regions correctly.

Each experiment was run as follows. In a sample of 1000 observations, we estimate $\beta = (\beta_1, \beta_2)$ using a probit model by maximum likelihood (ML) and the same probit model using the method suggested above which maximizes utility (i.e. set $h(X, \beta) = \Phi(\beta_1 + \beta_2 x)$) which we denote UM. We then draw an additional 1000 observations and examine how well, using the cutoff c from the utility maximization problem, each model forecasts the outcomes. The columns entitled KP are for known conditional probabilities used to forecast the outcomes. We report the breakdown of percentage of the observations classified into each correct and incorrect group as well as the total proportion correct and the utility level⁹ where the numbers are averaged over 10 replications of the experiment.

⁹Arbitrary rescalings of utility would allow the utility values to change, hence in the tables we provide

Table 1 reports the results for $c = 0.5$ for each of the distributions for X . In the first panel, we see that the UM estimates for β provide for a better classification and higher utility. The UM method results in correctly classifying the observations 73% as opposed to 58% when ML estimation is employed. Notice that the breakdown of which particular observations each classified correctly differs considerably between the methods. When $y = 1$, we have that the ML approach correctly classified the observation about one in every seven times, whereas for the UM method this was over three quarters of the time. In contrast when $y = -1$ we see that the ML estimates provided the correct classification about five times in every six, whereas the UM method obtained the correct answer for just under 4 of every six. Both methods did relatively well for the more prevalent (60% of the time) negative outcomes but the UM method did very well for the positive outcomes. The reason for the difference is evident from Figure 7. In this figure we plot $p(x)$ along with one of the draws of the estimated models for both UM and ML. Also included is a histogram of the covariate to show where the mass lies. Both estimators attempt to find index models that approximate $p(x)$ in the area where x has positive mass. Over the relevant range $p(x)$ rises steeply then falls, slowly flattening out. Since $p(x)$ starts low and ends high, both methods estimate upward sloping functions. However the ML method is much flatter, attempting to capture $p(x)$ in the right hand part of the distribution. This causes it to misclassify the observations for x around the hump — precisely the observations where we are more likely to see positive outcomes. The UM method by contrast uses the cutoff point and hence fits better around the best cutoff point for a single crossing model. So it ignores the second crossing (in a region where there are few observations) and gets the first crossing point correct. Hence it outperforms the probit 'density estimation' approach.

In the second panel the mass for the covariate is evenly spread across the entire x axis. Again, the UM method outperforms the ML method in classification. Here the proportion of the sample classified correctly into positives and negatives is very similar between the methods, however this masks large differences in exactly which types of observations are correctly specified. When $y = 1$ the UM method gets it right about half the time, whereas

the utility one would achieve if $p(x)$ were known as a benchmark.

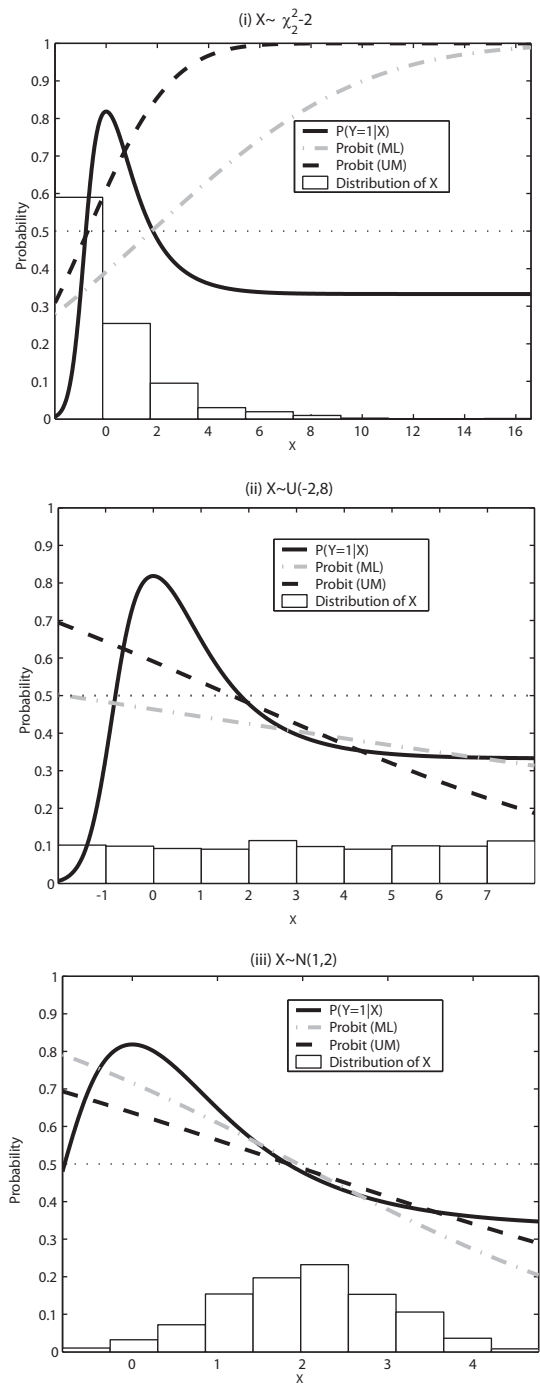


Figure 8: UM vs ML under a probit model specification for each design of X with $c = 0.5$

Table 1: Simulation results for $c(x) = 0.5$ and each design of X

Method	True +ve	False -ve	False +ve	True -ve	Correct	Average Utility
(i)						
UM	0.329	0.066	0.202	0.403	0.732	0.127
ML	0.056	0.338	0.082	0.524	0.580	-0.025
KP	0.275	0.120	0.116	0.490	0.764	0.159
(ii)						
UM	0.185	0.240	0.170	0.404	0.590	0.016
ML	0.017	0.409	0.062	0.513	0.530	-0.045
KP	0.184	0.242	0.082	0.493	0.676	0.102
(iii)						
UM	0.289	0.224	0.184	0.303	0.592	0.105
ML	0.325	0.188	0.217	0.270	0.595	0.108
KP	0.270	0.243	0.161	0.326	0.596	0.109

the ML method gets it correct about 10% of the time. Figure 8(ii) shows what is going on here. The ML method tries to approximate $p(x)$ over the full range of the graph, so effectively ignores the hump and estimates a downward sloping line that captures the general features except for the features when x is small. Because most of $p(x)$ is below the cutoff value it too is mostly under the cutoff value, and hence nearly always predicts $y = -1$. But for values of x near the hump, the majority of observations are positive and hence these are misclassified. Since it nearly always predicts $y = -1$ it does well when this is indeed true. The UM method, constrained to a single crossing, finds a crossing that leads to a good classification from a utility point of view. Hence it chooses to misclassify the first area where $p(x) < c$ but there are relatively few observations and hence relatively few classification errors. It picks the second crossing point well and hence gets it right (on average) at the area of the hump and also for large x .

Finally, for the third example for the distribution of the covariates the x 's are clustered around the second crossing point and there is no mass in the first region where $p(x) < c$.

Hence this is a single crossing problem with both models being constrained to have a single crossing of c . As such, both models correctly estimate the crossing point and both models perform very similarly. The UM model still outperforms slightly the ML method.

Table 2 examines these results when we change c , setting $c = 0.4$ and 0.6 respectively. The higher the cutoff the less strongly the utility function values matching the positives correctly. Hence for each estimator and each design for the covariate, the proportion of positive outcomes in the sample predicted falls. Since the number of positive values in the sample is the same within each design this means that the methods trade off losses through predicting positive outcomes less often with gains from predicting negative outcomes more often.

Note that the ML estimates for β do not change when we change the utility function. The mapping from these estimates to predictions of one or minus one does change, as the set of points at which $\Phi(x'_i \hat{\beta}_{ML}) > c$ changes. So the tradeoff as c rises depends on the estimates but the estimates do not use the information in the tradeoff. For the first design for X , since the ML (probit) estimate of $p(x)$ is upward sloping, this has the effect of increasing the value of x for which the method switches from predicting negative outcomes to positive outcomes. This effect can be seen in Figure 9(i). The optimal region over x for which a positive prediction should be made is exactly these moderate values for x , so the effect is not only to reduce the proportion of positive values but also strongly increase the number of false negatives. In each case though we can see that the method does not capture either of the points where $p(x)$ cuts c .

The UM method for estimation of β again does better in all cases in terms of utility maximization, as it should by construction. The estimates do change as we change c . For each case, as in the ML estimation method, the UM method forecasts a positive outcome for large x and a negative outcome for small x . However unlike the ML method as we increase c the estimates of the model parameters change so that the choice of this cutoff decreases as c increases. Indeed, at each value for c it is at the point where $p(x)$ first cuts c . Hence for each utility function it forecasts negative outcomes where x is small and $p(x) < c$ and positive outcomes over the 'hump' in $p(x)$. The method thus outperforms the ML method

Table 2: Simulation results for $c=0.4, 0.5, 0.6$ and each design of X .

Model	Method	True +ve	False -ve	False +ve	True -ve	Correct	Average Utility
(i)							
UM	$c=0.4$	0.351	0.049	0.220	0.379	0.730	0.204
	$c=0.5$	0.329	0.066	0.202	0.403	0.732	0.127
	$c=0.6$	0.224	0.175	0.135	0.466	0.690	0.021
ML	$c=0.4$	0.166	0.234	0.143	0.457	0.623	0.071
	$c=0.5$	0.056	0.338	0.082	0.524	0.580	-0.025
	$c=0.6$	0.017	0.382	0.036	0.565	0.582	-0.036
(ii)							
UM	$c=0.4$	0.352	0.061	0.421	0.167	0.518	0.071
	$c=0.5$	0.185	0.240	0.170	0.404	0.590	0.016
	$c=0.6$	0.000	0.422	0.000	0.578	0.578	0.000
ML	$c=0.4$	0.324	0.088	0.411	0.176	0.501	0.050
	$c=0.5$	0.017	0.409	0.062	0.513	0.530	-0.045
	$c=0.6$	0.000	0.422	0.000	0.578	0.578	0.000
(iii)							
UM	$c=0.4$	0.423	0.085	0.365	0.128	0.550	0.179
	$c=0.5$	0.289	0.224	0.184	0.303	0.592	0.105
	$c=0.6$	0.171	0.335	0.086	0.408	0.579	0.042
ML	$c=0.4$	0.445	0.062	0.391	0.102	0.547	0.185
	$c=0.5$	0.325	0.188	0.217	0.270	0.595	0.108
	$c=0.6$	0.150	0.356	0.065	0.429	0.579	0.052

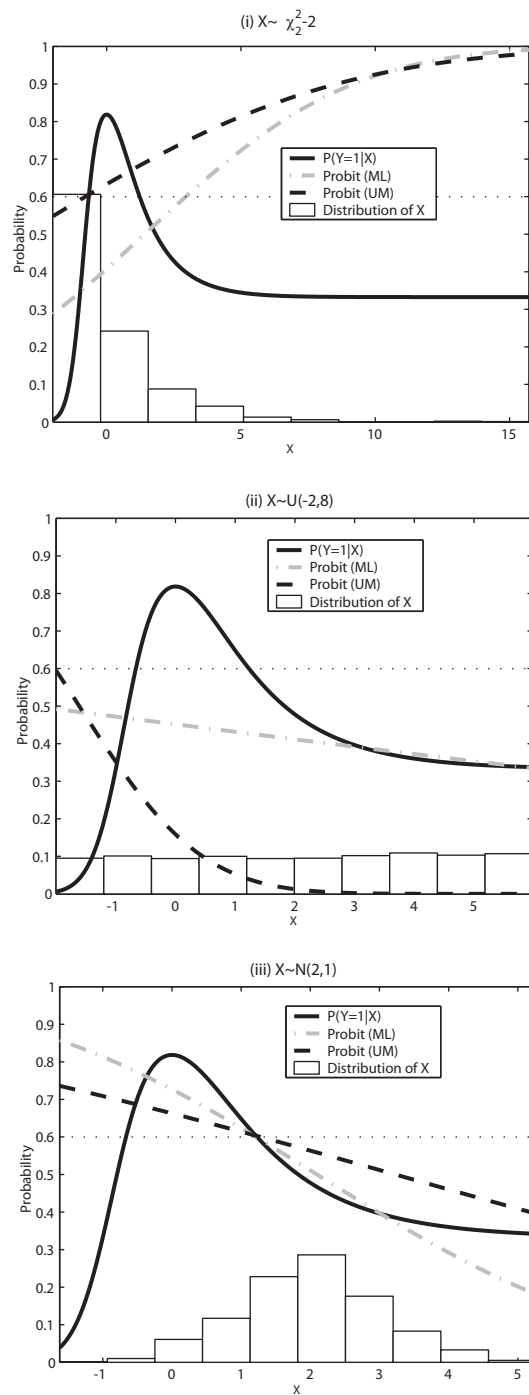


Figure 9: Asymptotic performance of UM vs ML when $c(x) = 0.6$

more strongly for utility functions that correspond to higher cutoffs.

For the second design the ML method estimates a downward sloping curve, as discussed above. The effects of changing the cutoff are pictured in Figure 9, where now a higher cutoff means that positive outcomes are predicted over a smaller range of x below this cutoff. The method thus predicts more negative outcomes over an increasing range where this is the optimal prediction. The UM estimator changes dramatically over the different cutoffs, including changing sign. As c gets large, both methods predict only negative outcomes.

We can also analyze in large samples the other methods discussed. We make two comparisons, the first based on a data generating process where $p(X)$ follows a logit specification, with $p(x) = \exp(x)/(1 + \exp(x))$ and X distributed as standard normal. The second accords to case (i) of the results above. We examine discriminant analysis (DA), where we impose that the covariance estimators are the same for each group and the groups are weighted evenly, the linear probability model (LPM), where we have estimated the linear coefficients by weighted least squares to take into account heteroskedasticity, a logit model, estimated by ML, and also a maximum utility model with a linear index. Results are contained in Table 3.

When $p(x)$ actually follows a logit specification, there is a single cutoff and so for all models there exists a set of parameters where they can provide optimal forecasts. In addition, since the logit model is correctly specified we expect that it will be able to classify as well as the UM method based on a linear model. This is indeed the case, where the differences amount to imperfections in the optimization procedure for the UM approach. Further, the discriminant analysis (DA) method and linear probability model (LPM) also do well, especially when the cutoff is at one half. However, this good performance is deceiving—costs further away from where the LPM and DA models cut $p(x)$ will cause these models to do quite poorly. They are essentially hit or miss, hence much of the similar performance in applications probably revolves around choosing loss functions that balance false positives and false negatives (i.e. the common choice of using one half as a cutoff). When the model of the conditional probability is misspecified, as in case (i), with the hump-shaped conditional probability function and the first design of X , the results are quite different. As with the

Table 3: Comparison of UM, logit, linear probability, and discriminant analysis

Method	Cutoff	True +ve	False -ve	False +ve	True -ve	Correct	Average Utility
Logit							
UM	c=0.4	0.400	0.091	0.257	0.252	0.651	0.229
	c=0.5	0.332	0.160	0.164	0.345	0.676	0.168
	c=0.6	0.232	0.264	0.081	0.424	0.656	0.111
ML (logit)	c=0.4	0.396	0.095	0.250	0.258	0.655	0.230
	c=0.5	0.334	0.157	0.163	0.346	0.680	0.171
	c=0.6	0.244	0.252	0.085	0.419	0.663	0.117
LPM	c=0.4	0.435	0.057	0.328	0.180	0.615	0.216
	c=0.5	0.333	0.159	0.162	0.347	0.680	0.171
	c=0.6	0.225	0.271	0.075	0.429	0.654	0.112
DA	c=0.4	0.320	0.172	0.156	0.353	0.672	0.216
	c=0.5	0.336	0.155	0.166	0.343	0.680	0.171
	c=0.6	0.340	0.156	0.163	0.341	0.681	0.097
(i)							
UM	c=0.4	0.351	0.049	0.220	0.379	0.731	0.205
	c=0.5	0.329	0.066	0.202	0.404	0.733	0.127
	c=0.6	0.224	0.175	0.135	0.466	0.689	0.021
ML (logit)	c=0.4	0.171	0.230	0.144	0.456	0.626	0.074
	c=0.5	0.062	0.333	0.085	0.521	0.582	-0.023
	c=0.6	0.020	0.379	0.038	0.563	0.583	-0.038
LPM	c=0.4	0.114	0.286	0.123	0.477	0.591	0.032
	c=0.5	0.025	0.369	0.048	0.558	0.583	-0.023
	c=0.6	0.005	0.394	0.013	0.588	0.593	-0.014
DA	c=0.4	0.218	0.177	0.153	0.452	0.670	0.064
	c=0.5	0.224	0.177	0.156	0.443	0.667	0.119
	c=0.6	0.219	0.180	0.156	0.445	0.664	-0.014

probit above, the logit now has a far inferior classification even with the correct cutoff. In all of the experiments we ran the logit and probit were almost indistinguishable in their predictive ability. The LPM also does not give great forecasts, although it is now different from the logit model (and slightly better in these experiments). The discriminant analysis method also works better than logit for this experiment, the good performance of this method when the cutoff is at 0.5 is due in part to the choice of the weighting on false positives and false negatives (even for all of these experiments).

We now examine the effects of adjusting the cost function so that it is not simply a constant everywhere. To this end we consider the experiment underlying case (i) and allow $c(x)$ to be a linear function of x . We choose the functions so that (approximately) they cut $p(x)$ at the second crossing point that would occur if $c = 0.5$. We consider four models, where $c(x)$ has slopes $-0.05, 0, 0.05$ and 0.1 . The cost functions are pictured in Figure 10. Recall that the higher is the cutoff function the relatively more valuable are correct predictions of the negative outcome. The negatively sloped $c(x)$ thus puts a greater weight relative to the constant $c = 0.5$ on predicting negative outcomes at lower values for x and a smaller weight on predicting negatives at higher values. The positively sloped cutoff functions have the opposite effect, the greater the slope the more useful it will be from a utility perspective to predict the high frequency of positive values that occurs near the 'hump' in $p(x)$.

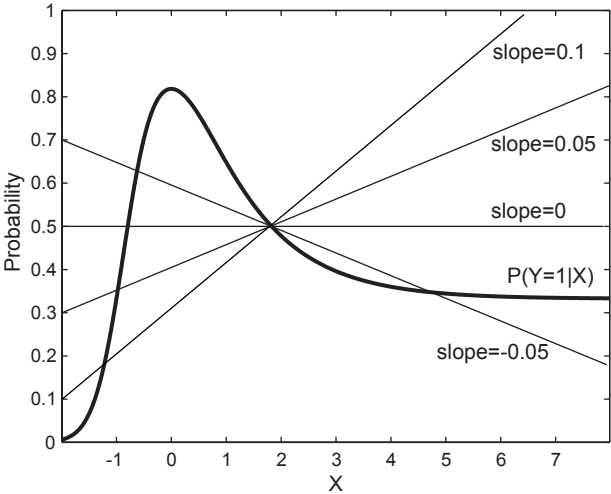


Figure 10: $c(x)$ with various slopes

Table 4 shows results for the probit estimation and the linear maximum utility approach. Each entry is for a single replication (single Monte Carlo run with 1000 observations in the evaluation sample). Recall again that the estimated ML model is exactly as in the first panel in Figure 8 above, as they are not affected by the change in the cost function. What is affected is the mapping of these to positive and negative outcomes. The UM method adjusts to best use the information in the cost function conditional on the specified linear functional form. For $c(x)$ negatively sloped and constant the classification for the probit model does not change, because $p(x) > c(x)$ for exactly the same values for x . What does change is how these count—with the negative slope the common negative outcomes at very low x , predicted this way by the probit, have a greater value and hence utility rises. Utility also goes up for the UM method.

Table 4: UM vs ML when $c(x)$ a linear function in x (case (i))

Slope		True +ve	False -ve	False +ve	True -ve	Correct	Average Utility
-0.05	UM	0.324	0.067	0.158	0.451	0.775	0.109
	ML	0.061	0.330	0.067	0.542	0.603	0.015
0	UM	0.334	0.057	0.175	0.434	0.768	0.159
	ML	0.061	0.330	0.067	0.542	0.603	-0.006
0.05	UM	0.343	0.048	0.192	0.417	0.760	0.251
	ML	0.055	0.336	0.060	0.549	0.604	-0.027
0.10	UM	0.368	0.023	0.261	0.348	0.716	0.436
	ML	0.329	0.062	0.542	0.067	0.396	0.425

For a slight positive slope to $c(x)$ we see that the probit method alters the classification slightly. This is due to the relative slopes of the probit estimated model and the cutoff function, which results in a slightly smaller range over which positive predictions are made. The loss function now weights correct positive predictions more highly relative to incorrect positive values. Since the ML method mainly predicts negative values over this range, it fares poorly in terms of utility. The UM method reacts to the now relatively higher weight on getting the positive predictions correct for smaller values by predicting more positive

values and we see a large increase in utility. Finally, when the cutoff function $c(x)$ is steep enough, it is below the ML model for small values of x and hence the classification changes dramatically—the ML method now predicts many more positive outcomes. The UM model also reacts to this. Overall we see that the ML method, by virtue of ignoring the utility function in the estimation, can perform very poorly even when both models are similarly misspecified.

6 Conclusion

It is often the case that some binary decision—a choice of yes or no—need be made in the presence of uncertainty. In the economics literature this can be credit granting choices, predicting bankruptcy, predicting the sign of asset price movements, predicting credit card fraud, predicting economic crises etc. In each of these cases there is uncertainty, i.e. whether or not the loan will be paid back, whether or not the firm actually goes bankrupt. The same problem arises in program design problems such as college admission, job programs etc. where the value of admission depends on how well the program is used by the individual, i.e. do they complete the program successfully, do they get a job afterwards?

Many popular methods can be thought of as conducting the two step approach of first estimating the conditional forecast density—here the probability that the outcome is “successful”—and then basing the decision on this density estimate. The usual justification for the two step approach is that knowledge of the conditional probability of success as a function of the observed covariates is sufficient for any decision maker to solve their optimization problem. However the fact that the conditional density has to be estimated raises at least two concerns. First, is it reasonable to ignore the decision maker’s loss function in estimation? Second, many of the applications of this approach then evaluate the success or lack thereof of the procedure through the proportion of a sample which corresponds to correct predictions. This then is a situation of estimation under one loss function and evaluation under another. If this strategy is to provide good results then it is more through luck than design.

This paper considers both design and estimation of models that directly incorporate

the utility function. In the binary decision problem the utility function takes a convenient simple form, where the function to be optimized is an extension of the types of maximum score functions analyzed by Manski (1975, 1985). The extension arises through placing this method within a utility maximization framework. Econometrically the extensions amount to (i) the presence of variables that affect utility entering as weights on the scores, and (ii) extending beyond linear functions of the data in specifying the scores. The first of these extensions is motivated by the utility setup, there is in principle no reason why utility should be the same over all situations. The second arises through the need for flexible functions to adequately capture the richness of the classification. We provide analytic properties of the estimators.

Finally, we show that the density forecast approach is as suggested by its construction essentially a “hit or miss” affair. The second step in this two step approach is to use the utility function to provide a cutoff for which conditional probabilities which are above this cutoff become forecasts of a success. Thus success of the method relies on estimating well the conditional probability at points around this cutoff, and otherwise being on the correct side of the cutoff. However in misspecified models we show that this will not necessarily be the case, and that the properties of this approach can be far inferior to the method we suggest. Namely, we show that instead of estimating the conditional probability function first, and finding the cutoff second, it is in general a better strategy to estimate the best decision/forecast/allocation directly, already making use of the decision maker’s utility function in the estimation procedure. In essence the estimator we suggest abstracts from the unnecessary fitting of the conditional probability in regions where doing so is not useful in providing useful forecasts.

References

- [1] ALTMAN, E. (1968): Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance* 23: 589-609.
- [2] AMEMIYA, T. (1981): Qualitative Response Models—A Survey. *Journal of Economic Literature* 19: 1483-1536.
- [3] ANDREWS, D.W.K. (1987): Consistency in Nonlinear Econometric Models: A Generic Uniform Law of Large Numbers. *Econometrica* 55: 1465-1471.
- [4] ANDREWS, D.W.K. (1994): Empirical Process Methods in Econometrics. In: *Handbook of Econometrics*, vol. IV, eds. R.F. Engle and D.L. McFadden. Elsevier.
- [5] 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.
- [6] BOYES, W., D. HOFFMAN AND S. LOW (1989): An Econometric Analysis of the Bank Credit Scoring Problem. *Journal of Econometrics* 40: 3-14.
- [7] CAUDILL, S.B. (2003): Predicting Discrete Outcomes with the Maximum Score Estimator: the case of the NCAA Men's Basketball Tournament. *International Journal of Forecasting* 19: 313-317.
- [8] CRAMER, J.S. (1998): Predictive Performance of the Binary Logit Model in Unbalanced Samples, manuscript.
- [9] DIEBOLD, F.X., T.A. GUNTHER AND A.S. TAY (1998): Evaluating Density Forecasts With Applications to Financial Risk Management. *International Economic Review* 39: 863-883.
- [10] DIMITRAS, A.I., S.H. ZANAKIS AND C. ZOPOUNIDAS (1996): A survey of business failures with an emphasis on prediction methods and industrial applications. *European Journal of Operations Research* 90: 487-513.

- [11] FISHER, R.A. (1936): The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics* 7: 179-88.
- [12] GRANGER, C.W.J. (1969): Prediction with a Generalized Cost of Error Function. *Operational Research* 20: 199-207.
- [13] GRANGER, C.W.J. AND M.H. PESARAN (2000): Economic and Statistical Measures of Forecast Accuracy. *Journal of Forecasting* 19: 537-560.
- [14] HALL, P. AND C. HEYDE (1980): *Martingale Limit Theory and Its Application*. Academic Press: New York.
- [15] HOROWITZ, J.L. (1992): A Smoothed Maximum Score Estimator for the Binary Response Model. *Econometrica* 60: 505-531.
- [16] IBRAGIMOV, I.A. AND YU.V. LINNIK (1971): *Independent and Stationary Dependent Sequences of Random variables*. Woltors-Noordhoff: Groningen.
- [17] KIM, J. AND D. POLLARD (1990): Cube Root Asymptotics. *Annals of Statistics* 18: 191-219.
- [18] LEUNG, M.T., H. DAOUK AND A-S. CHEN (2000): Forecasting Stock Indices: A Comparison of Classification and Level Estimation Models. *International Journal of Forecasting* 16: 173-190.
- [19] MADDALA, G.S. (1983): *Limited Dependent and Qualitative Variables in Econometrics*. Econometric Society Monograph No. 3. Cambridge University Press.
- [20] MANSKI, C.F. (1975): Maximum Score Estimation of the Stochastic Utility Model of Choice. *Journal of Econometrics* 3: 205-228.
- [21] MANSKI, C.F. (1985): Semiparametric Analysis of Discrete Response. Asymptotic Properties of the Maximum Score Estimator. *Journal of Econometrics* 27: 313-333.
- [22] MANSKI, C.F. AND T.S. THOMPSON (1989): Estimation of Best Predictors of Binary Response. *Journal of Econometrics* 40: 97-123.

- [23] MARTIN, D. (1977): Early warning of Bank Failure: A Logit Regression Approach. *Journal of Banking and Finance* 1: 249-276.
- [24] MIN, Q. AND S. YANG (2003): Forecasting consumer credit card adoption: What can we learn about the utility function? *International Journal of Forecasting* 71-83.
- [25] OHLSEN, J. (1980): Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research* 18: 109-131.
- [26] PESARAN, M.H. AND S. SKOURAS (2001): Decision-Based Methods for Forecast Evaluation. In: M.P. Clemens and D.F. Hendry (eds.) *Companion to Economic Forecasting*, Basil Blackwell.
- [27] POLLARD, D. (1984): *Convergence of Stochastic Processes*. Springer-Verlag, New York.
- [28] POWELL, J. (1994): Estimation of Semiparametric Models. In: *Handbook of Econometrics* Vol. 4, Engle, R. and D. McFadden, eds. North-Holland, Amsterdam.
- [29] QI, M. AND S. YANG (2003): Forecasting Consumer Credit Card Adoption: What Can We Learn About the Utility Function? *International Journal of Forecasting* 71-83.
- [30] RESNICK, S.I. (1999): *A Probability Path*. Birkhäuser: Boston.
- [31] SRINIVASAN, V. AND Y.H. KIM (1987): Credit Granting: A Comparative Analysis of Classification Procedures. *Journal of Finance* 17: 665-681.
- [32] VAN DER VAART, A.W. AND J.A. WELLNER (1996): *Weak Convergence and Empirical Processes*. Springer Verlag: New York.
- [33] WANG, Y. AND I.H. WITTEN (2002): Modeling for Optimal Probability Prediction, monograph, University of Waikato.
- [34] WEISS, A.A. (1986): Estimating Time Series Models Using the Relevant Cost Function. *Journal of Applied Econometrics* 11: 539-560.
- [35] WHITE, H. (2000): *Asymptotic Theory for Econometricians*. Academic Press: Orlando, Fla.

- [36] WILSON, R.L. AND R. SHANDA (1994): Bankruptcy Prediction using Neural Networks. Decision Support Sciences 11: 431-447.

Appendix: Proofs

Notation Throughout the appendix we will use the following notation:

$$s(y, x, \theta) = b(x)[y + 1 - 2c(x)]\text{sign}[h(x, \theta)].$$

Hence,

$$S(\theta) = E_{Y,X}[s(Y, X, \theta)] \quad \text{and} \quad S_n(\theta) = n^{-1} \sum_{i=1}^n s(Y_i, X_i, \theta).$$

Lemma 1 *If Condition 1 and Condition 2 parts (b) and (c) hold then the function $\theta \mapsto S(\theta)$ is continuous at all $\theta \in \Theta$.*

Proof Conditions 1 and 2(b) ensure that $S(\theta)$ is well defined. Fix any $\theta \in \Theta$ and consider an arbitrary sequence $\theta_m \subset \Theta$ such that $\theta_m \rightarrow \theta$. Continuity follows from showing that $S(\theta_m) \rightarrow S(\theta)$ as $m \rightarrow \infty$. Let $F_\theta^+ = \{\omega : h(X, \theta) > 0\}$. If $\omega \in F_\theta^+$, then by $\theta_m \rightarrow \theta$ and the continuity of $h(\cdot, \theta)$ (the first part of Condition 2(c)), we can find an integer M such that $h(X, \theta_m) > 0$ for all $m > M$. Thus for such an m

$$\begin{aligned} s(Y, X, \theta_m) &= b(X)\{Y + 1 - 2c(X)\}\text{sign}[h(X, \theta_m)] \\ &= b(X)\{Y + 1 - 2c(X)\}\text{sign}[h(X, \theta)] \\ &= s(Y, X, \theta). \end{aligned}$$

Defining $F_\theta^- = \{\omega : h(X, \theta) < 0\}$ and choosing $\omega \in F_\theta^-$ leads by the same argument to the above result holding. By the second part of Condition 2(d) we have $P(F_\theta^+ \cup F_\theta^-) = 1$ so we have that $s(Y, X, \theta_m) \xrightarrow{a.s.} s(Y, X, \theta)$ where the exception occurs on θ but not θ_m .

Since $|s(Y, X, \theta_m)| \leq 2b(X)$ and $b(X)$ is integrable by Condition 1 then application of the dominated convergence theorem results in

$$S(\theta_m) = E[s(Y, X, \theta_m)] \rightarrow E[s(Y, X, \theta)] = S(\theta).$$

as $m \rightarrow \infty$. ■

Proof of Proposition 1 Existence of $\max_{\theta \in \Theta} S(\theta)$ follows from continuity of the objective function in θ and compactness of Θ . Continuity follows from Lemma 1 given Condition 1 and Condition 2(b) and (c). Compactness is Condition 2(a). ■

Lemma 2 Let (X, d) be a metric space and let $f : D \rightarrow \mathbb{R}$ and $f_n : D \rightarrow \mathbb{R}$, $n = 1, 2, \dots$, be functions defined on the set $D \subset X$. Let $M \subset D$ denote the set of maximizers of f over D and $M_n \subset D$ the set of maximizers of f_n over D . Suppose

(i) M is nonempty.

(ii) M_n is nonempty for each $n = 1, 2, \dots$

(iii) f_n converges uniformly to f on D , i.e. $\lim_{n \rightarrow \infty} \sup_{x \in D} |f_n(x) - f(x)| = 0$.

Let $x^* \in M$ and let $x_n^* \in M_n$. Then

$$(a) f(x_n^*) \rightarrow f(x^*) \quad \text{and} \quad (b) f_n(x_n^*) \rightarrow f(x^*).$$

Proof Since x^* maximizes $f(x)$,

$$\begin{aligned} 0 &\leq f(x^*) - f(x_n^*) \\ &= f(x^*) - f_n(x^*) + f_n(x^*) - f(x_n^*) \\ &\leq f(x^*) - f_n(x^*) + f_n(x_n^*) - f(x_n^*) \end{aligned}$$

where the last line follows as x_n^* maximizes $f_n(x)$. Hence

$$\begin{aligned} |f(x_n^*) - f(x^*)| &\leq |f(x^*) - f_n(x^*) + f_n(x_n^*) - f(x_n^*)| \\ &\leq |f(x^*) - f_n(x^*)| + |f_n(x_n^*) - f(x_n^*)| \\ &\leq 2 \sup_{x \in D} |f_n(x) - f(x)| \end{aligned}$$

which goes to zero by assumption (iii). Also

$$\begin{aligned} |f_n(x_n^*) - f(x^*)| &= |f_n(x_n^*) - f(x_n^*) + f(x_n^*) - f(x^*)| \\ &\leq |f_n(x_n^*) - f(x_n^*)| + |f(x_n^*) - f(x^*)| \\ &\leq 3 \sup_{x \in D} |f_n(x) - f(x)| \end{aligned}$$

and hence this also goes to zero by assumption (iii). ■

Proof of Proposition 2 We first show that $S_n(\theta) \xrightarrow{a.s.} S(\theta)$ uniformly in θ . This follows through application of the results in Andrews (1987). We will show that assumptions A1, A2 and A6 of Andrews (1987) hold for the result to be shown. Assumption A1 is that the parameter space be compact, which is assumed directly (Condition 2(a)).

Let $B(\theta_0, \epsilon)$ denote an open ball with radius $\epsilon > 0$ centered on $\theta_0 \in \Theta$ and suppose that the random variables $\{(Y_i, X_i')\}_{i=1}^\infty$ are defined on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (the assumption of completeness is without loss of generality). Andrews' A2(a) requires that

$$\left\{ \sup_{\theta \in B(\theta_0, \epsilon) \cap \Theta} s(Y_i, X_i, \theta) \right\} \quad (10)$$

be a sequence of random variables, i.e. $\mathcal{F}/\mathcal{B}(\mathbb{R})$ -measurable functions, for all $\theta_0 \in \Theta$, $\epsilon > 0$. This follows from Appendix C of Pollard (1984) given the measurability of $s(\cdot, \cdot, \cdot)$ which follows in turn from Condition 1 and Condition 2 part (b), the fact that $B(\theta_0, \epsilon) \cap \Theta$ is a Borel subset of \mathbb{R}^p for all $\epsilon > 0$, $\theta_0 \in \Theta$ and that (Y_i, X_i') is defined on a complete probability space. Now for A2(b),

$$\left| \sup_{\theta \in B(\theta_0, \epsilon) \cap \Theta} s(Y_i, X_i, \theta) \right| = |b(X_i)[Y_i + 1 - 2c(X_i)]| \leq 2b(X_i),$$

which by Condition 1 part (b) ensures integrability. This, along with Condition 2 part (d) and Thm. 3.35 of White (2000), allows the application of the ‘‘pointwise’’ LLN for stationary ergodic sequences (Thm. 3.34, White 2000) to (10). The same holds for the infimum, which obtains A2.

To show the first part of A6 note that the almost sure continuity of the mapping $\theta \mapsto s(Y_i, X_i, \theta)$ was established in the proof of Lemma 1 and required Condition 1 and Condition 2 parts (b) and (c). Part (b) of A6 follows as

$$\begin{aligned} E \sup_{i \geq 1, \theta \in \Theta} |s(Y_i, X_i, \theta)| &= E \sup_{\theta \in \Theta} |s(Y_1, X_1, \theta)| \\ &\leq 2Eb(X_1) \end{aligned}$$

which is finite. The first equality follows from Condition 2 part (d).

Extending this result to showing that $S_n(\hat{\theta}_n) \xrightarrow{a.s.} S(\theta^*)$ involves applying Lemma 2 above where we have that $\hat{\theta}_n$ maximizes $S_n(\theta)$ and θ^* maximizes $S(\theta)$. ■

Lemma 3 *Let (X, d) be a metric space and let $f : D \rightarrow \mathbb{R}$ and $f_n : D \rightarrow \mathbb{R}$, $n = 1, 2, \dots$, be functions defined on the set $D \subset X$. Let $M \subset D$ denote the set of maximizers of f over D and $M_n \subset D$ the set of maximizers of f_n over D . Suppose*

- (i) *D is compact and f is continuous (so that M is nonempty).*

(ii) M_n is nonempty for each $n = 1, 2, \dots$

(iii) f_n converges uniformly to f on D , i.e. $\lim_{n \rightarrow \infty} \sup_{x \in D} |f_n(x) - f(x)| = 0$.

Let $x_n^* \in M_n$ and let $x_{n_j}^*$ be a convergent subsequence of x_n^* . Then $x_{n_j}^* \rightarrow x^*$ for some $x^* \in M$ as $j \rightarrow \infty$. Furthermore, $d(x_n^*, M) \rightarrow 0$.

Proof Let $f^* = \max_{x \in D} f(x)$. As $x_n^* \in D$, a compact set, x_n^* has a convergent subsequence $x_{n_j}^*$. Suppose the first claim of the lemma is false, i.e. $x^\circ \equiv \lim_{j \rightarrow \infty} x_{n_j}^*$ is not contained in M . By D compact, $x^\circ \in D$. Since $x^\circ \notin M$, $f(x^\circ) < f^*$. Let $\epsilon = [f^* - f(x^\circ)]/2 > 0$. By the continuity of f at x° , there exists $\delta > 0$ such that

$$f(x^\circ) - \epsilon < f(x) < f(x^\circ) + \epsilon = f^* - \epsilon \quad \forall x \in B(x^\circ, \delta) \cap D,$$

where $B(x^\circ, \delta)$ denotes the open ball with radius δ centered on x° . Since $x_{n_j}^* \rightarrow x^\circ$, there exists an integer N such that $x_{n_j}^* \in B(x^\circ, \delta)$ for all $j > N$, implying, from above,

$$f(x_{n_j}^*) < f^* - \epsilon \quad \forall j > N. \quad (11)$$

Let f_{n_j} denote the subsequence of f_n corresponding to x_{n_j} . By condition (iii) (uniform convergence), there exists an integer K such that $\sup_{x \in D} |f_{n_j}(x) - f(x)| < \epsilon/2$ for all $j > K$. In particular,

$$f_{n_j}(x_{n_j}^*) < f(x_{n_j}^*) + \epsilon/2 \quad \text{and} \quad (12)$$

$$f_{n_j}(x^*) > f(x^*) - \epsilon/2 = f^* - \epsilon/2 \quad \forall x^* \in M \quad \forall j > K. \quad (13)$$

For $j > \max\{N, K\}$ we can combine (12), (11) and (13) to obtain

$$f_{n_j}(x_{n_j}^*) < f(x_{n_j}^*) + \epsilon/2 < (f^* - \epsilon) + \epsilon/2 = f^* - \epsilon/2 < f_{n_j}(x^*).$$

This contradicts the fact that $x_{n_j}^*$ is a maximizer of f_{n_j} and hence proves the first assertion of the lemma.

To prove the second assertion, suppose the claim is not true. Then for some $\epsilon > 0$ there is a subsequence $x_{n_k}^*$ such that $d(x_{n_k}^*, M) > \epsilon$ for all k . Since $x_{n_k}^*$ is contained in the compact set D , it must have a convergent subsequence $x_{n_{k(j)}}^*$. By the assertion just proven above, the limit point of $x_{n_{k(j)}}^*$, as $j \rightarrow \infty$, is contained in M . But this is impossible given the way $x_{n_k}^*$ was selected. This contradiction proves the second assertion. ■

Proof of Proposition 3 Apply Lemma 3 with $D = \Theta$, $f = S$, $f_n = S_n$, $x^* = \theta^*$ and $x_n^* = \hat{\theta}_n$. The compactness of Θ is directly assumed, continuity of $S(\theta)$ is shown by Lemma 1 and $\max_{\theta} S_n(\theta)$ always exists. Almost sure uniform convergence of $S_n(\theta)$ to $S(\theta)$ is shown in the proof of Proposition 2 above. ■

Lemma 4 *Suppose Condition 3 is satisfied. Then there exist constants $A > 0$ and $\lambda > 0$ such that for all $r > 0$ sufficiently small*

$$\sup_{\theta_0 \in \Theta} E \left\{ \sup_{\theta \in B_d(\theta_0, r) \cap \Theta} \left| s(Y, X, \theta) - s(Y, X, \theta_0) \right|^2 \right\} \leq Ar^\lambda,$$

where λ is determined by Condition 3 part (c')(i) and $B_d(\theta, r)$ denotes the open ball with radius r centered on θ w.r.t. the Euclidian metric d .

Corollary 1 *Suppose $\Theta^* = \arg \max_{\Theta} S(\theta)$ consists of maximizers of a single type. Under Condition 3, $d(\hat{\theta}_n, \Theta^*) \rightarrow_{a.s.} 0$ implies $\rho(\hat{\theta}_n, \theta^*) \rightarrow_{a.s.} 0$ for any $\theta^* \in \Theta^*$.*

Proof of Lemma 4 Fix $\theta_0 \in \Theta$ and $r > 0$. To avoid clutter, we will simply write “ \sup_{θ} ” instead of “ $\sup_{\theta \in B_d(\theta_0, r) \cap \Theta}$ ” in the following derivations.

$$\begin{aligned} & E \left\{ \sup_{\theta} |s(X, Y, \theta) - s(X, Y, \theta_0)|^2 \right\} \\ &= E \left\{ |b(X)[Y + 1 - c(X)]|^2 \sup_{\theta} |\text{sign}[h(X, \theta)] - \text{sign}[h(X, \theta_0)]|^2 \right\} \\ &\leq C^2 \times E \left\{ \sup_{\theta} |\text{sign}[h(X, \theta)] - \text{sign}[h(X, \theta_0)]|^2 \right\} \\ &= 4C^2 \times P \left\{ \sup_{\theta} |\text{sign}[h(X, \theta)] - \text{sign}[h(X, \theta_0)]| = 2 \right\} \\ &= 4C^2 \times P \left\{ \exists \theta \in B_d(\theta_0, r) \cap \Theta \text{ such that } \text{sign}[h(X, \theta)] \neq \text{sign}[h(X, \theta_0)] \right\} \\ &= 4C^2 \times P \left\{ \bigcup_{\theta \in B_d(\theta_0, r) \cap \Theta} \{ \text{sign}[h(X, \theta)] \neq \text{sign}[h(X, \theta_0)] \} \right\}, \end{aligned} \tag{14}$$

where on the third line use is made of the fact that $|b(X)[Y + 1 - c(X)]| \leq C$ for some constant C (Condition 1 part (b)). Now, by Condition 3 part (c')(i) (Lipschitz), there exists a constant $\lambda > 0$ and a function $L(x) \geq 0$ such that

$$h(X, \theta_0) - L(X)r^\lambda \leq h(X, \theta) \leq h(X, \theta_0) + L(X)r^\lambda \quad \forall \theta \in B(\theta_0, r) \cap \Theta.$$

Therefore, if θ is within r of θ_0 , then $\text{sign}[h(X, \theta)]$ must agree with $\text{sign}[h(X, \theta_0)]$ on the event

$$\{\omega : 0 < h(X, \theta_0) - L(X)r^\lambda \text{ or } h(X, \theta_0) + L(X)r^\lambda \leq 0\}.$$

On the other hand, $\text{sign}[h(X, \theta)]$ is potentially different from $\text{sign}[h(X, \theta_0)]$ on the complement event

$$\{\omega : h(X, \theta_0) - L(X)r^\lambda \leq 0 < h(X, \theta_0) + L(X)r^\lambda\}.$$

Notice that this event does not depend on θ . Thus,

$$\begin{aligned} & \bigcup_{\theta \in B(\theta_0, r)} \{\omega : \text{sign}[h(X, \theta)] \neq \text{sign}[h(X, \theta_0)]\} \\ & \subset \{\omega : h(X, \theta_0) - L(X)r^\lambda \leq 0 < h(X, \theta_0) + L(X)r^\lambda\}. \end{aligned}$$

Therefore, we can continue the string of inequalities under (14) by writing

$$\begin{aligned} \dots & \leq 4C^2 \times P[h(X, \theta_0) - L(X)r^\lambda \leq 0 < h(X, \theta_0) + L(X)r^\lambda] \\ & = 4C^2 \times P[-r^\lambda < h(X, \theta_0)/L(X) \leq r^\lambda]. \end{aligned}$$

By Condition 3 part (c')(ii), the distribution of the random variable $h(X, \theta_0)/L(X)$ is absolutely continuous w.r.t. Lebesgue measure conditional on, say, the first $k-1$ components of X . Denote this density by $f_{\theta_0}(\cdot | x^{(1)}, \dots, x^{(k-1)})$. It is also assumed that we can choose $r > 0$ sufficiently small such that $f_{\theta_0}(z | x^{(1)}, \dots, x^{(k-1)}) \leq M$ for some $M > 0$ whenever $|z| \leq r^\lambda$, and both M and r are independent of $x^{(1)}, \dots, x^{(k-1)}$ and θ_0 . Therefore,

$$\begin{aligned} & P[-r^\lambda < h(X, \theta_0)/L(X) \leq r^\lambda] \\ & = E\{P[-r^\lambda < h(X, \theta_0)/L(X) \leq r^\lambda | X^{(1)}, \dots, X^{(k-1)}]\} \\ & = E \int_{-r^\lambda}^{r^\lambda} f_{\theta_0}(z | X^{(1)}, \dots, X^{(k-1)}) dz \leq 2r^\lambda M. \end{aligned}$$

Combining this result with previous inequalities yields

$$E \left\{ \sup_{\theta \in B(\theta_0, r) \cap \Theta} |s(X, Y, \theta) - s(X, Y, \theta_0)|^2 \right\} \leq 8C^2 M r^\lambda$$

for all $r > 0$ sufficiently small. Since the bound on the RHS is independent of the choice of $\theta_0 \in \Theta_0$, the claim follows. ■

Proof of Corollary 1 Since Θ^* is a closed set under Condition 3(c'), there exist $\theta_n^* \in \Theta^*$ such that $d(\hat{\theta}_n, \Theta^*) = d(\hat{\theta}_n, \theta_n^*)$ for each n . Using these θ_n^* , we can write, 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 ρ -distance between any two elements of Θ^* is zero (in fact, equality must hold above). Let $r_n = d(\hat{\theta}_n, \theta_n^*)$. Since $r_n \rightarrow_{a.s.} 0$, by Lemma 4 we can choose n large enough so that

$$\rho(\hat{\theta}_n, \theta_n^*) \leq \left\{ E \left[\sup_{\theta \in B_d(\theta_n^*, r_n + 1/n)} |s(Y, X, \theta) - s(Y, X, \theta_n^*)|^2 \right] \right\}^{1/2} \leq A^{1/2} (r_n + 1/n)^{\lambda/2}$$

for some $A > 0$ and $\lambda > 0$. Letting n go to infinity completes the proof. ■

Proof of Proposition 4 We complete the argument in the main text (see p. 23) by showing that under Condition 3 the function

$$J_n(\theta) = n^{1/2}[S_n(\theta) - S(\theta)] = n^{-1/2} \sum_{i=1}^n \{s(Y_i, X_i, \theta) - E[s(Y, X, \theta)]\}$$

is stochastically equicontinuous w.r.t. the semimetric ρ . That is, for any $\epsilon, \eta > 0$ there must exist $\delta > 0$ such that¹⁰

$$\limsup_{n \rightarrow \infty} P \left(\sup_{\rho(\theta_1, \theta_2) < \delta} |J_n(\theta_1) - J_n(\theta_2)| > \eta \right) < \epsilon. \quad (15)$$

For a detailed discussion of stochastic equicontinuity see van der Vaart and Wellner (1996, Ch. 1.5), Andrews (1994), Andrews and Pollard (1994).

We establish (15) by checking the assumptions of Theorem 2.2. of Andrews and Pollard (1994), where the family $\{s(\cdot, \cdot, \theta) : \theta \in \Theta\}$ plays the role of \mathcal{F} and $\{(Y_i, X_i)\}$ plays the role of $\{\xi_i\}$. The use of the theorem is justified by noting that $J_n(\theta)$ can be regarded as an empirical process indexed by the class of functions $\{s(\cdot, \cdot, \theta) : \theta \in \Theta\}$.

- (*The condition on the mixing coefficients*) This assumption is directly included in Condition 3 part (d').

¹⁰The measure theoretic details ensuring that the event under (15) has a well defined probability are not discussed here. The interested reader is referred to Pollard (1984, Appendix C). Alternatively, outer probability could be used to guard against measurability problems.

- (*The integral condition on the bracketing numbers*) By the discussion preceding Thm. 2.2 of *ibid.*, Lemma 4 ensures that for $x > 0$, the bracketing numbers $N(x)$ of the family $\{s(\cdot, \cdot, \theta) : \theta \in \Theta\}$ are bounded by $Kx^{-2p/\lambda}$, where K is some positive constant, λ is determined by Condition 3 part (c')(i) and p is the dimension of Θ . It is easy to see then that the integral condition on the bracketing numbers is satisfied if $Q \geq 2$ and $\gamma > 0$ is chosen such that $Q/(2 + \gamma) > p/\lambda$, which is assumed directly in Condition 3 part (d').

Finally, we note that the class of functions $\{s(\cdot, \cdot, \theta) : \theta \in \Theta\}$ is uniformly bounded by Condition 1. Hence, the theorem applies and the stochastic equicontinuity of $J_n(\theta)$ follows.

■

Proof of Proposition 5 Write

$$\begin{aligned}
n^{1/2}[M_n(\hat{\theta}_{2,n}) - M_n(\hat{\theta}_{i,n})] &= n^{1/2}[S_n(\hat{\theta}_{2,n}) - p_2g(n)] - n^{1/2}[S_n(\hat{\theta}_{i,n}) - p_i g(n)] \\
&= n^{1/2}[S_n(\hat{\theta}_{2,n}) - S(\theta_2^*)] - n^{1/2}[S_n(\hat{\theta}_{i,n}) - S(\theta_i^*)] \\
&\quad + n^{1/2}[S(\theta_2^*) - S(\theta_i^*)] + (p_i - p_2)n^{1/2}g(n) \\
&\equiv Z_{i,n} + n^{1/2}[S(\theta_2^*) - S(\theta_i^*)] + (p_i - p_2)n^{1/2}g(n),
\end{aligned}$$

where the definition of $Z_{i,n}$, $i = 1, 3$ is apparent. When $Z_{i,n}$ is $O_p(1)$ (as under Prop. 4), for any $\epsilon > 0$ there exists $\Delta > 0$ and $N \in \mathbb{N}$ such that

$$P[\Delta \geq Z_n \geq -\Delta] > 1 - \epsilon \text{ for all } n > N. \quad (16)$$

Model 1 (underfit) vs. model 2 (pseudo true model). We have

$$P[M_n(\hat{\theta}_{2,n}) > M_n(\hat{\theta}_{1,n})] = P\{Z_n > n^{1/2}[S(\theta_1^*) - S(\theta_2^*) + (p_2 - p_1)g(n)]\}.$$

As $S(\theta_1^*) - S(\theta_2^*) < 0$, $p_2 - p_1 > 0$ and $g(n) = C \log(n)/\sqrt{n} \rightarrow 0$, the RHS of the inequality inside the curly brackets tends to minus infinity with n . Given $\epsilon > 0$, find $\Delta > 0$ and $N \in \mathbb{N}$ such that (16) holds. We can also find $K \in \mathbb{N}$ such that $n^{1/2}[S(\theta_2^*) - S(\theta_1^*) + (p_2 - p_1)g(n)] < -\Delta$ for all $n > K$. Thus, for $n > \max\{N, K\}$,

$$P\{Z_n > n^{1/2}[S(\theta_1^*) - S(\theta_2^*) + (p_2 - p_1)g(n)]\} \geq P[Z_n \geq -\Delta] \geq P[\Delta \geq Z_n \geq -\Delta] > 1 - \epsilon.$$

Model 3 (overfit) vs. model 2 (pseudo true model). In this case $S(\theta_2^*) = S(\theta_3^*)$, so

$$P[M_n(\hat{\theta}_{2,n}) > M_n(\hat{\theta}_{3,n})] = P\{Z_n > (p_2 - p_3)n^{1/2}g(n)\}.$$

As $p_2 - p_3 < 0$ and $n^{1/2}g(n) = C \log(n) \rightarrow \infty$, the RHS of the inequality in the curly brackets once again tends to minus infinity with n . By the same argument as above, the probability in question converges to one. ■