Horowitz, Joel; Lee, Sokbae

Working Paper

Binary classification with the maximum score model and linear programming

cemmap working paper, No. CWP16/25

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Horowitz, Joel; Lee, Sokbae (2025) : Binary classification with the maximum score model and linear programming, cemmap working paper, No. CWP16/25, Centre for Microdata Methods and Practice (cemmap), The Institute for Fiscal Studies (IFS), London, https://doi.org/10.47004/wp.cem.2025.1625

This Version is available at:
https://hdl.handle.net/10419/324454

cemmap
centre for microdata methods and practice

Binary classification with the maximum score model and linear programming

Joel L. Horowitz
Sokbae Lee

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP16/25

Economic and Social Research Council

BINARY CLASSIFICATION WITH THE MAXIMUM SCORE MODEL AND LINEAR PROGRAMMING

JOEL L. HOROWITZ[1] AND SOKBAE LEE[2] [3]

ABSTRACT. This paper presents a computationally efficient method for binary classification using Manski's (1975,1985) maximum score model when covariates are discretely distributed and parameters are partially but not point identified. We establish conditions under which it is minimax optimal to allow for either non-classification or random classification and derive finite-sample and asymptotic lower bounds on the probability of correct classification. We also describe an extension of our method to continuous covariates. Our approach avoids the computational difficulty of maximum score estimation by reformulating the problem as two linear programs. Compared to parametric and nonparametric methods, our method balances extrapolation ability with minimal distributional assumptions. Monte Carlo simulations and empirical applications demonstrate its effectiveness and practical relevance.

KEYWORDS: Binary classification; maximum score estimation; partial identification; finite-sample and asymptotic inference; extrapolation

1. INTRODUCTION

This paper introduces a computationally efficient method for applying Manski's (1975, 1985) maximum score model of binary response to binary classification when covariates may be discretely distributed and parameters are partially but not point identified. We establish conditions under which it is minimax optimal to allow for either non-classification or random classification and derive finite-sample and asymptotic lower bounds on the probability of correct classification.

[1]DEPARTMENT OF ECONOMICS, NORTHWESTERN UNIVERSITY, EVANSTON, IL 60208, USA.
[2]CENTRE FOR MICRODATA METHODS AND PRACTICE, INSTITUTE FOR FISCAL STUDIES, 7 RIDGMOUNT STREET, LONDON, WC1E 7AE, UK.
[3]DEPARTMENT OF ECONOMICS, COLUMBIA UNIVERSITY, NEW YORK, NY 10027, USA.
E-mail addresses: joel-horowitz@northwestern.edu, sl3841@columbia.edu.
Date: July 25, 2025.
We would like to thank Chuck Manski, Adam Rosen, Jörg Stoye, and Elie Tamer for their helpful comments. Part of this research was conducted while we were visiting Cemmap and the Department of Economics at University College London.

In Manski's model, the binary dependent variable Y is related to a finite-dimensional vector of explanatory variables X as follows:

$$Y = I\left(\beta' X - U \geq 0\right), \quad \mathbb{P}(U \leq 0 \mid X) = \tau, \tag{1.1}$$

where $I(\cdot)$ is the indicator function, β is an unknown vector of constants, and $0 < \tau < 1$, implying that zero is the τ-quantile of U conditional on X. The distribution of U is otherwise unrestricted and may depend on X in an unknown way (i.e., heteroskedasticity of unknown structure). Given an independent random sample $\{(Y_i, X_i) : i = 1, \ldots, n\}$, the maximum score estimator of β is defined as:

$$\hat{\beta} = \arg\max_{b \in \mathbf{B}} \sum_{i=1}^{n} Y_i I\left(b' X_i > 0\right), \tag{1.2}$$

where $\mathbf{B}$ is the parameter space.

If β were known, an individual with covariate X_* would be classified as $Y = 1$ if $\beta' X_* > 0$ and $Y = 0$ otherwise. However, when X follows a discrete distribution, β and $\beta' X_*$ are not point identified. Under mild conditions, $\beta' X_*$ is contained in an identified interval, denoted here by $[c_L, c_H]$. In this paper, we propose two simple classification rules. In one rule, an individual with covariate X_* is classified as $Y = 1$ if $c_L > 0$, as $Y = 0$ if $c_H < 0$, and remains unclassified if $c_L \leq 0 \leq c_H$. We provide a decision-theoretic framework in which this classification rule is shown to be minimax optimal. This minimax optimal classification rule relies on two elements: a lower cost associated with abstention (non-classification) compared to misclassification, and the explicit availability of non-classification as an option. In the second classification rule, abstention is not possible, and decisions are made via randomization. In this case, we propose a modified decision rule: classify as $Y = 1$ if $c_L > 0$, as $Y = 0$ if $c_H < 0$, and assign $Y = 1$ or $Y = 0$ randomly with equal probability when $c_L \leq 0 \leq c_H$. We again establish conditions under which this randomized classification rule is minimax optimal.

We further present estimators for $[c_L, c_H]$ and construct both finite-sample and asymptotic confidence intervals, denoted by $[\hat{c}_L, \hat{c}_H]$. The classification rule follows the same logic: we classify an individual as $Y = 1$ if $\hat{c}_L > 0$ and as $Y = 0$ if $\hat{c}_H < 0$, with non-classification or random classification occurring when $\hat{c}_L \leq 0 \leq \hat{c}_H$. We say that an individual is classified correctly if (s)he would have the same classification if c_L and c_U were known. Theoretically, we provide upper bounds on the probability of misclassification and, thereby, lower bounds on the probability of correct classification.

Computing $\hat{\beta}$ (solving the optimization problem (1.2)) is a notoriously difficult, NP-hard combinatorial optimization problem. In addition, $\hat{\beta}$ converges in probability to β at the slow rate of $n^{-1/3}$ when β is point identified. The objective of this paper, however, is estimation of the interval $[c_L, c_H]$ that contains $\beta' X_*$, not point identification or estimation of β. We give conditions under which $[c_L, c_H]$ can be estimated by solving two linear programming problems, one for c_L and one for c_H. Computationally efficient algorithms and software for solving linear programming problems are widely available, so we avoid the computational difficulty of solving (1.2).

While the underlying model involves the parameter vector β, our primary interest lies in conducting binary classification rather than in precisely estimating β itself. Although the bounds for β may be wide, informative classification remains possible. By focusing on the signs of linear combinations of covariates rather than on point estimates of coefficients, we leverage the structure of the identified set to assign classifications reliably. Thus, even when parameter uncertainty is substantial, the classification rule can yield meaningful and robust decisions.

Various methods exist for binary classification. Fully nonparametric methods, such as estimating $\mathbb{P}(Y = 1 \mid X = X_*)$ nonparametrically and classifying based on whether this probability exceeds τ, make minimal assumptions but suffer from poor extrapolation ability. If the covariates follow a discrete distribution, these methods fail to classify individuals whose X_* is not observed in the sample and are imprecise if there are few observations of X_*. Parametric methods, including logit, probit, and discriminant analysis, allow extrapolation but rely on restrictive distributional assumptions. The method we propose balances these trade-offs: it permits extrapolation while being less restrictive than fully parametric approaches and some semiparametric models. For example, semiparametric single-index models satisfying the quantile restriction are special cases of the maximum score model.

The remainder of this paper is structured as follows. Section 2 reviews the literature on maximum score estimation. Section 3 formally describes the model. Section 4 provides a decision-theoretic framework for the proposed classification rule. Section 5 outlines our estimation procedure and discusses finite-sample and asymptotic inference with discretely distributed covariates. It also presents lower bounds on the probability of correct classification. Finite-sample inference typically yields conservative bounds in moderate-sized samples common in economics applications, as it imposes

only weak constraints on the distribution of the underlying random variables and ensures validity for the worst-case scenario under those constraints. In contrast, asymptotic inference yields tighter bounds but does not guarantee validity in finite samples. Section 6 details our computational algorithms. Section 7 extends the framework to accommodate continuous covariates, specifically incorporating Manski and Tamer (2002)-type bounds when a continuous vector is treated as observed only within a hypercube. Section 8 provides empirical illustrations: Section 8.1 presents an empirical example using data from Kessler, Low, and Sullivan (2019); Section 8.2 illustrates the extension discussed in Section 7, again using the same data; Section 8.3 presents another empirical application using data from Corno, Hildebrandt, and Voena (2020). Section 9 reports a Monte Carlo experiment demonstrating the effectiveness of our classification rules. Section 10 concludes with a discussion of key findings and potential directions for future research. Appendix A contains the proofs of the theoretical results presented in the main text. Appendix B presents Monte Carlo simulations comparing finite-sample and asymptotic inference methods. Finally, Appendix C extends the framework to accommodate clustered dependence and survey sampling weights.

2. Literature Review

The literature on the maximum score model is concerned mainly with identification and estimation of β, methods for computing $\hat{\beta}$ in (1.2), and inference. The literature is large, and we cite only a limited set of papers that treat these topics.

2.1. **Identification, Asymptotic Properties, and Related Methods.**

Manski (1988) provided classical results for point identification. Subsequently, advances have been made in terms of partial identification: see, e.g., Manski and Tamer (2002); Komarova (2013); Blevins (2015); Chen and Lee (2019); and Khan, Komarova, and Nekipelov (2024) among others.

When β is point identified, $\hat{\beta}$ converges in probability to β at the rate $n^{-1/3}$, and $n^{1/3}(\hat{\beta} - \beta)$ has an intractable asymptotic distribution (see, e.g., Kim and Pollard, 1990; Seo and Otsu, 2018). For inference, subsampling and bootstrap methods are considered in e.g., Delgado, Rodrıguez-Poo, and Wolf (2001); Abrevaya and Huang (2005); Lee and Pun (2006); Patra, Seijo, and Sen (2015); and Cattaneo, Jansson, and Nagasawa (2020) among others; however, they are computationally demanding because it is required to solve (1.2) repeatedly.

Horowitz (1992, 2002) developed smoothed maximum score estimation to mitigate the difficulties coming from the maximum score objective function. The smoothed maximum score estimator is asymptotically normal with a faster rate of convergence and enables the bootstrap to provide asymptotic refinements. Alternative approaches include: finite sample inference (Rosen and Ura, 2025); Bayesian methods (Benoit and Van den Poel, 2012; Walker, 2024); non-Bayesian Laplace type approaches (Jun, Pinkse, and Wan, 2015, 2017); and two-stage methods (Gao, Xu, and Xu, 2022) among many others.

2.2. **Computational Challenges.** One major challenge in applying maximum score estimation lies in the computational complexity of the objective function. The maximum score objective is a piecewise constant function, with a finite range of discrete values. Consequently, while this function always reaches a maximum, the maximizer is not necessarily unique, making computation of the maximum score estimates NP-hard (see, e.g., Johnson and Preparata, 1978). Early algorithms for addressing maximum score estimation were developed by Manski and Thompson (1986) and Pinkse (1993). Second-generation approaches leverage advances in mixed integer programming (MIP) solvers, along with significant increases in computing power since the development of the first-generation methods. Florios and Skouras (2008) presented strong numerical evidence that an MIP-based method outperforms earlier techniques. Similarly, Kitagawa and Tetenov (2018) employed an alternative MIP formulation to solve a maximum score-type problem for treatment choice rules, where they maximized an empirical welfare criterion resembling the maximum score function. Chen and Lee (2018, 2020) extended this approach to ℓ_0-constrained and ℓ_0-penalized empirical risk minimization for high-dimensional binary classification. However, the MIP approach faces intrinsic scalability limitations. To the best of our knowledge, numerical studies applying MIP to maximum score estimation typically remain confined to datasets with a modest sample size (e.g., around 1,000) and a limited number of parameters. In contrast, our approach is based on linear programming, which is considerably more scalable and capable of handling larger datasets with a greater number of parameters.

3. Model

This section describes the model we treat and the parameters of interest. We begin with definitions and notation.

Let $X \in \mathbb{R}^q$ be a random or non-stochastic (fixed) vector. If X is random, assume for now that it is discretely distributed with J mass points $\{x_j : j = 1, .., J\}$ and probability mass function $p(x_j) > 0$. The case of continuously distributed components of X is treated in Section 7. If X is fixed, let it have J possible values $\{x_j : j = 1, .., J\}$. In this case, we treat a design in which there are n observations of (Y, X) and $n_j > 0$ of these observations have $X = x_j$. Define $p(x_j) = n_j/n$, the fraction of observations in the design for which $X = x_j$. For any random event A, let $P(A \mid X = x_j)$ denote the probability of A conditional on $X = x_j$ if X is random and the probability of A at $X = x_j$ if X is fixed.

The model is

$$Y = \begin{cases} 1 & \text{if } \beta' X - U \geq 0 \\ 0 & \text{if } \beta' X - U < 0 \end{cases},$$

where U is an unobserved random variable that satisfies $\mathbb{P}(U \leq 0 \mid X = x_j) = \tau$ for all $j = 1, \ldots, J$, some constant $0 < \tau < 1$, and some unknown vector of constant parameters $\beta \in \mathbb{R}^q$. We assume that this model is correctly specified unless stated otherwise. Note that the $(1 - \tau)$ quantile of Y conditional on $X = x_j$ is

$$\mathbb{Q}_{1-\tau}(Y \mid X = x_j) = I(\beta' x_j \geq 0), \tag{3.1}$$

where $I(\cdot)$ is the usual indicator function. A scale normalization is needed for identification of β. We use $\beta_1 = 1$, where β_1 is the first component of β.

For each $j = 1, \ldots, J$ define

$$f(x_j) \equiv [\mathbb{P}(Y = 1 | X = x_j) - \tau] \text{ and } d_j \equiv \text{sgn}[f(x_j)],$$

where for a real number u, $\text{sgn}(u) = 1$ if $u > 0$, $\text{sgn}(u) = -1$ if $u < 0$, and $\text{sgn}(u) = 0$ if $u = 0$. Let $r'\beta$, where $r \in \mathbb{R}^q$ is a non-zero vector of constants, be a linear combination of components of β. The parameter β is not point identified when X is discretely distributed with finitely many mass points or fixed with finitely many possible values. The linear combination $r'\beta$ is not point identified if at least one component of $(r_2, \ldots, r_J)'$ is non-zero, but it is interval identified. Let $[c_L, c_U]$ be the identified interval.

3.1. **Assumptions.** We assume the following regularity conditions.

Assumption 1. *One of the following conditions holds:*

(a) *(i) X is fixed and $p(x_j) \equiv n_j/n > 0$ is known. (ii) For each $j = 1, ..., J$, $\{Y_{\ell,j} : \ell = 1, ..., n_j\}$ is an independent random sample from $\mathbb{P}(Y|X = x_j)$.*

(b) *(i) X is random. (ii) $\{(Y_i, X_i) : i = 1, \ldots, n\}$ is an independent random sample from the joint distribution of (Y, X). (iii) For all $j = 1, \ldots, J$, $p(x_j) \equiv \mathbb{P}(X = x_j) > 0$.*

Assumption 2. *$|f(x_j)| > 0$ for all $j = 1, \ldots, J$.*

Assumption 3. *$\beta \in \mathbf{B} \subset \mathbb{R}^q$, where*

$$\mathbf{B} \equiv \{(1, b_2, \ldots, b_q) : (b_2, \ldots, b_q) \text{ belongs to a known compact set } \}.$$

Under Assumption 2, there is an $\varepsilon > 0$ such that $\min_{1 \leq j \leq J} |\beta' x_j| = \varepsilon$. Under Assumption 3, the parameter space for β is a known compact set. Define

$$g_0(x_j) = \begin{cases} \mathbb{E}(Y_{\ell,j} - \tau) p(x_j) & \text{if } X \text{ is fixed} \\ \mathbb{E}\left[(Y - \tau) I(X = x_j)\right] & \text{if } X \text{ is random} \end{cases}.$$

If X is random, we include the indicator function $I(X = x_j)$ in $g_0(x_j)$ so that it can be estimated as a sample average without a random denominator.

Remark 1. Assumption 2 excludes boundary cases where $f(x_j) = 0$ (equivalently, $x_j'\beta = 0$) for some j. Although this simplifies identification, it is not without cost. Cases with $f(x_j) = 0$ yield equality constraints ($x_j'\beta = 0$), which provide stronger identifying power and could shrink the identified interval for $r'\beta$. Relaxing Assumption 2, however, introduces significant practical challenges. In empirical applications, the true conditional probability $\mathbb{P}(Y = 1 \mid X = x_j)$ is typically unknown and must be estimated. Even if the true probability equals exactly τ, estimators rarely match this precisely, necessitating an approximation interval $[\tau - \ell_n, \tau + \ell_n]$ for some tuning parameter ℓ_n. Determining an appropriate ℓ_n and adjusting estimation and inference methods accordingly is nontrivial. Moreover, existing finite-sample inference results suggest that exact equality constraints contribute little to inference about β (see Rosen and Ura, 2025, Appendix G). Thus, despite their theoretical appeal, equality constraints may offer limited practical benefit. Given these considerations, we maintain Assumption 2 for clarity and practicality, but acknowledge that exploring ways to incorporate or approximate these boundary cases remains an important avenue for future research.

3.2. **Partial Identification.** In model (1.1), $Y = 1$ if and only if $\beta' X \geq U$, where $\mathbb{P}(U \leq 0 \mid X) = \tau$. Under Assumption 2, we have

$$\mathbb{P}(Y = 1 \mid X = x) > \tau \Leftrightarrow \beta' x > 0,$$
$$\mathbb{P}(Y = 1 \mid X = x) < \tau \Leftrightarrow \beta' x < 0.$$

Therefore, under Assumption 2, the identified set for β is

$$\{b \in \mathbf{B} : d_j = \operatorname{sgn}(b'x_j); b_1 = 1, |b'x_j| \geq \varepsilon; j = 1, \ldots, J\}.$$

This is equivalent to

$$d_j (x_{j1} + b_2 x_{j2} + \ldots + b_q x_{jq}) \geq \varepsilon \quad (j = 1, \ldots, J).$$

The identified upper and lower bounds, c_U and c_L, on $r'\beta$ are optimal values of the objective functions of the linear programs

$$\underset{b \in \mathbf{B}}{\text{maximize}} \left(\underset{b \in \mathbf{B}}{\text{minimize}} \right) : \ r'b \tag{3.2}$$

subject to

$$d_j (x_{j,1} + b_2 x_{j,2} + \cdots + b_q x_{j,q}) \geq \varepsilon \tag{3.3}$$

for all $j = 1, \ldots, J$.

Remark 2. The linear programs (3.2)-(3.3) assume that ε is known. If ε is unknown and replaced by a quantity that is too large, the feasible region (3.3) is too small and does not contain the entire identified interval for $r'\beta$. If ε is replaced by a smaller quantity, the feasible region is larger but is not tight. In what follows, we set $\varepsilon = 0$. The resulting interval obtained by solving (3.2)-(3.3) contains the interval $[c_L, c_U]$ but is not tight.

Remark 3. Assumption 2 is needed because $d_j b'x_j = 0$ does not imply that $d_j = 0$ and $b'x_j = 0$. See Remark 1 for further discussion of Assumption 2.

Under Assumption 1, $d_j \equiv \operatorname{sgn}[f(x_j)]$ and $g_{0,j} \equiv g_0(x_j)$ have the same sign. Therefore, the linear programs (3.2)-(3.3) with $\varepsilon = 0$ are equivalent to

$$\underset{b \in \mathbf{B}}{\text{maximize}} \left(\underset{b \in \mathbf{B}}{\text{minimize}} \right) : \ r'b \tag{3.4}$$

subject to

$$g_{0,j} (x_{j,1} + b_2 x_{j,2} + \cdots + b_q x_{j,q}) \geq 0 \tag{3.5}$$

for all $j = 1, \ldots, J$.

4. A Decision-Theoretic Framework

In this section, we present a decision-theoretic framework for our classification rules. We start with the case of a known, point-identified β. Let $\hat{y} \in \{0, 1\}$ denote a binary

classifier. For covariate values x within the support of X, the function

$$f(x) = \mathbb{P}(Y = 1 \mid X = x) - \tau$$

is nonparametrically identified. At the population level, the optimal classification rule assigns

$$\hat{y} = \begin{cases} 1, & \text{if } f(x) > 0, \\ 0, & \text{if } f(x) < 0. \end{cases} \tag{4.1}$$

However, when x lies outside the support of X, the function $f(x)$ is not identified, and classification based on (4.1) is not feasible. Classification based on $f(x)$, which is what we rely on when β is unknown and only partially identified, does not extend to off-support covariate values. In contrast, if β is known, we can replace (4.1) with the equivalent linear rule

$$\hat{y} = \begin{cases} 1, & \text{if } \beta' x > 0, \\ 0, & \text{if } \beta' x < 0. \end{cases} \tag{4.2}$$

This formulation enables classification at any covariate value x, including those not in the support of X.

Remark 4 (Classification under Point Identification of β). The classification rules in (4.1) and (4.2) can be interpreted as optimal decision rules minimizing the $(1-\tau)$-absolute loss function, as we have $\mathbb{Q}_{1-\tau}(Y \mid X = x) = I(\beta' x \geq 0)$ (recall (3.1)). This result is formally established as Corollary 4 in Manski (1988), who also emphasizes the role of extrapolation beyond the support of X.

We now move to the case of partially identified β. Although $\beta' x$ is only *interval-identified*:

$$\beta' x \in [c_L(x),\, c_U(x)],$$

this interval provides information about the sign of $f(x)$, which can be leveraged to extrapolate $f(x)$ and classify off-support values of x. To account for the partial identification of $\beta' x$, we consider two scenarios: one that includes the option of non-classification, and another that allows for random classification.

4.1. **Adding an Option of Non-Classification.**

In this subsection, we extend the classification options from $\{0, 1\}$ to $\{0, 1, \emptyset\}$, where $\emptyset$ denotes "non-classification."

Consider the following loss function:

$$L(\hat{y}, y) = \begin{cases} C_b, & \text{if } \hat{y} = 1 \text{ and } f(x) > 0, \\ C_b, & \text{if } \hat{y} = 0 \text{ and } f(x) < 0, \\ C, & \text{if } \hat{y} = 0 \text{ and } f(x) > 0, \\ C, & \text{if } \hat{y} = 1 \text{ and } f(x) < 0, \\ C_\emptyset, & \text{if } \hat{y} = \emptyset \text{ and } f(x) > 0, \\ C_\emptyset, & \text{if } \hat{y} = \emptyset \text{ and } f(x) < 0, \end{cases} \tag{4.3}$$

where $0 \leq C_b < C_\emptyset < C < \infty$.

The loss function $L(\hat{y}, y)$ reflects the relative costs associated with three possible classification actions: assigning 1, assigning 0, or abstaining (denoted by $\emptyset$). Correct classifications incur a baseline loss C_b, interpreted as the cost of making a decision even with oracle knowledge of $f(x)$; this cost need not be zero, as the realized outcome Y can differ from the classification implied by $f(x)$. Incorrect classifications, where $\hat{y}$ disagrees with the sign of $f(x)$, incur a higher loss C, reflecting the penalty for wrong decisions such as incorrect treatments or verdicts. Abstention incurs an intermediate loss $C_\emptyset$, satisfying $C_b < C_\emptyset < C$, and models the idea that refraining from classification is preferable to misclassification but costlier than a correct decision. Abstention seems natural when partial identification of $f(x)$ limits prediction for off-support x. The exact values of C_b, $C_\emptyset$, and C are immaterial for our classification rule, provided the ordering $0 \leq C_b < C_\emptyset < C < \infty$ holds. Overall, the structure of the loss function encourages a cautious classification strategy. The following theorem formalizes this intuition by characterizing the decision rule that minimizes the worst-case loss under this loss structure.

Theorem 4.1 (Optimal Classification Rule with Option of Abstention)**.** *Let* $[c_L(x), c_U(x)]$ *be the identified lower and upper bounds on* $\beta' x$*. Consider the loss function defined in* (4.3)*, where* $0 \leq C_b < C_\emptyset < C < \infty$*. Then the minimax optimal classification rule (i.e., the rule minimizing the worst-case loss, shortened as OCR) is:*

$$\text{OCR}(x) \;=\; \begin{cases} 1, & \textit{if } c_L(x) > 0, \\ 0, & \textit{if } c_U(x) < 0, \\ \emptyset, & \textit{otherwise (i.e., if } 0 \in [c_L(x),\, c_U(x)]), \end{cases}$$

where $\emptyset$ *refers to "non-classification."*

In words, we adopt the following classification rule: predict 1 if the entire identified region for $\beta' x$ is strictly positive; predict 0 if it is strictly negative; and abstain otherwise.

When x is off-support, we determine a bounding interval $[c_L(x), c_U(x)]$ for $\beta' x$. If this interval is entirely positive, we predict 1; if entirely negative, we predict 0; otherwise, we abstain to avoid incurring the high misclassification cost C. The key insight is that $C_\emptyset$ is strictly smaller than C, making non-classification optimal in ambiguous cases.

Remark 5 (Role of Cost Structure and Non-Classification Option). Two components are crucial to our framework: (i) the cost structure $C_\emptyset < C$, and (ii) the inclusion of non-classification (denoted by $\emptyset$) in the choice set. Regarding the first, if instead $C_\emptyset > C$ holds, it will never be optimal to select $\emptyset$ even when $0 \in [c_L(x), c_U(x)]$, implying that our proposed rule is no longer minimax optimal. Regarding the second, Theorem 4.1 applies specifically to settings that permit non-classification. Section 4.2 describes a decision rule for situations in which classification as either 0 or 1 is mandatory for all individuals.

Remark 6 (Relation to Selective Classification). There is a branch of the machine learning literature known as selective classification or classification with a reject option (e.g., El-Yaniv and Wiener, 2010). In this framework, a classification rule typically consists of two components: (i) a selection function that determines whether to classify, and (ii) a binary classifier that assigns a label when classification occurs. Although the frameworks share surface-level similarities, the objectives underlying our classification rule are distinct. For a recent review of selective classification and related topics, see Hendrickx, Perini, Van der Plas, Meert, and Davis (2024). In the economics literature, Breza, Chandrasekhar, and Viviano (2025) recently proposed a method for policy learning that incorporates an abstention option.

Remark 7 (Dempster–Shafer). Dempster–Shafer (DS) theory (Dempster, 2008) extends classical probability by introducing a third category, "don't know", alongside "true" and "false". It assigns a triplet (p, q, r) of non-negative values summing to one, representing evidence for, against, and uncertainty about an assertion, respectively. In short, DS theory allows $r > 0$ to reflect ambiguity. While related in spirit, our notion of non-classification remains grounded in conventional probability.

4.2. **Random Classification.** One might argue against the option of abstention when the identification interval includes zero. If one truly faces a binary choice

problem, abstention may not be possible. As an alternative, one might consider a randomized choice between the two options. Indeed, there is a precedent for randomization as a sensible approach to making binary decisions under ambiguity (Manski, 2009).

Since abstention is not feasible in this subsection, we modify the loss function in (4.3) as follows:

$$L(\hat{y}, y) = \begin{cases} C_b, & \text{if } \hat{y} = 1 \text{ and } f(x) > 0, \\ C_b, & \text{if } \hat{y} = 0 \text{ and } f(x) < 0, \\ C, & \text{if } \hat{y} = 0 \text{ and } f(x) > 0, \\ C, & \text{if } \hat{y} = 1 \text{ and } f(x) < 0, \end{cases} \tag{4.4}$$

where $0 \leq C_b < C < \infty$.

Suppose we randomly classify an observation as 1 with probability p and as 0 with probability $1 - p$. The expected loss is then:

$$L = \begin{cases} pC_b + (1-p)C, & \text{if } \beta' x > 0, \\ (1-p)C_b + pC, & \text{if } \beta' x < 0. \end{cases} \tag{4.5}$$

To minimize the expected loss, it is necessary to know the true sign of $\beta' x$. However, since this is unknown, we instead aim to minimize the maximum regret. The regret is defined as the expected loss minus the oracle loss C_b:

$$\text{Regret} = \begin{cases} (1-p)(C - C_b), & \text{if } \beta' x > 0, \\ p(C - C_b), & \text{if } \beta' x < 0. \end{cases} \tag{4.6}$$

Theorem 4.2 (Optimal Classification Rule with an Option of Random Classification)**.** *Let $[c_L(x), c_U(x)]$ be the identified lower and upper bounds on $\beta' x$. Consider the loss function defined in* (4.4)*, where $0 \leq C_b < C < \infty$. Then, the minimax-regret optimal classification rule (i.e., the rule that minimizes the worst-case expected loss relative to the oracle case of knowing the true sign of $f(x)$, shortened as MMR) is:*

$$\text{MMR}(x) \;=\; \begin{cases} 1, & \textit{if } c_L(x) > 0, \\ 0, & \textit{if } c_U(x) < 0, \\ RC, & \textit{otherwise (i.e., if } 0 \in [c_L(x),\, c_U(x)]), \end{cases}$$

where RC refers to classifying as 1 or 0 with equal probability.

Theorem 4.2, building on the framework of Manski (2009), provides a robust rationale for assigning classifications randomly with equal probability when the identification interval includes zero. Although the theorem is stated in terms of minimizing maximum regret, it also characterizes the minimax-optimal rule under expected loss. This equivalence holds in our setting because the oracle loss C_b remains constant regardless of the true sign of $\beta' x$.

5. Estimation and Inference

Section 5.1 describes estimation of c_L and c_U. Sections 5.2 and 5.3 describe our method of inference about $[c_L, c_U]$. Section 5.4 gives the lower bounds on the probability of correct classification when c_L and c_U are replaced by our estimates.

5.1. **Estimation.** In applications, $g_{0,j}$ is unknown. We estimate it by one of two methods, depending on whether X is fixed or random. If X is fixed, estimate $g_{0,j}$ by

$$\hat{g}\left(x_j\right) = n_j^{-1} \sum_{\ell=1}^{n_j} \left(Y_{\ell,j} - \tau\right) p(x_j),$$

If X is random, estimate $g_{0,j}$ by

$$\hat{g}(x_j) = n^{-1} \sum_{i=1}^{n} \left(Y_i - \tau\right) I\left(X_i = x_j\right).$$

Both estimators are unbiased:

$$\mathbb{E}\left[\hat{g}\left(x_j\right)\right] = g_0\left(x_j\right); \quad j = 1, \ldots, J.$$

Now let $\mathcal{G}_\alpha$ be a confidence region for $g_0 \equiv (g_{0,1}, \ldots, g_{0,J})$ satisfying

$$P(\mathcal{G}_\alpha) \equiv \mathbb{P}\left\{g_0 \equiv (g_{0,1}, \ldots, g_{0,J}) \in \mathcal{G}_\alpha\right\} \geq 1 - \alpha \tag{5.1}$$

for some $0 < \alpha < 1$. Here, $\mathcal{G}_\alpha$ may be a finite-sample or asymptotic confidence region. Section 5.2 explains how to construct finite-sample confidence regions based on the two versions of $\hat{g}(x_j)$. Section 5.3 explains how to construct asymptotic confidence regions.

The estimators of c_U and c_L, denoted respectively by $\hat{c}_U$ and $\hat{c}_L$, are the optimal values of the objective functions of the following optimization problems:

$$\underset{b \in \mathbf{B};\, g \in \mathcal{G}_\alpha}{\text{maximize}} \left(\underset{b \in \mathbf{B};\, g \in \mathcal{G}_\alpha}{\text{minimize}} \right) : \; r'b \tag{5.2}$$

subject to

$$g_j\left(x_{j,1}+b_2x_{j,2}+\cdots+b_qx_{j,q}\right)\geq 0 \tag{5.3}$$

for all $j=1,\ldots,J$.

The following theorem, which is similar to Theorem 2.1 of Horowitz and Lee (2023), is the basis for our inference and classification.

Theorem 5.1. *Let* (5.1) *hold. Then,*

$$\mathbb{P}\left[\hat{c}_L\leq c_L\leq r'\beta\leq c_U\leq\hat{c}_U\right]\geq\mathbb{P}\left\{g\in\mathcal{G}_\alpha\right\}\geq 1-\alpha.$$

Theorem 5.1 shows that if (5.1) holds, the estimated interval $[\hat{c}_L,\hat{c}_U]$ contains the population interval $[c_L,c_U]$ with finite-sample or asymptotic probability at least $(1-\alpha)$, depending on whether $\mathcal{G}_\alpha$ is a finite-sample or asymptotic confidence region. Theorem 5.1 implies that to obtain a confidence region for $[c_L,c_U]$, it suffices to obtain the confidence region $\mathcal{G}_\alpha$. Sections 5.2 and 5.3 explain how to construct finite-sample and asymptotic confidence regions $\mathcal{G}_\alpha$ that satisfy (5.1).

In terms of classification, Theorem 5.1 establishes lower bounds on the probability of correct classification when the estimated bounds allow classification (i.e., when $\hat{c}_L>0$ or $\hat{c}_U<0$). However, it does not provide guarantees for cases where $\hat{c}_L\leq 0\leq\hat{c}_U$, meaning classification is indeterminate. Sections 5.4 and 5.5 provide bounds that include the possibility of indeterminate classification. Nevertheless, as the sample size n increases, the probability of correct classification for all cases approaches 1, since $\hat{c}_L$ and $\hat{c}_U$ consistently estimate c_L and c_U. This result aligns with the standard intuition behind size and power comparisons in hypothesis testing.

5.2. **Finite Sample Inference.** This section obtains finite-sample confidence regions $\mathcal{G}_\alpha$ for fixed and random X.

5.2.1. *Fixed Design.* Let X be fixed and Assumption 1(a) hold. Then

$$-\tau p\left(x_j\right)\leq\left(Y_{\ell,j}-\tau\right)p(x_j)\leq(1-\tau)p\left(x_j\right) \tag{5.4}$$

because $Y=0$ or 1. An application of Hoeffding's inequality yields

$$\mathbb{P}\left[\left|\hat{g}\left(x_j\right)-g_0\left(x_j\right)\right|\geq p\left(x_j\right)t_j\right]\leq 2\exp\left(-2n_jt_j^2\right)$$

for any $t_j>0$ and each $j=1,\ldots,J$. Moreover,

$$\hat{g}\left(x_j\right)-p\left(x_j\right)t_j\leq g_0\left(x_j\right)\leq\hat{g}\left(x_j\right)+p\left(x_j\right)t_j$$

for all $j = 1, \ldots, J$ with probability at least $1 - 2\sum_{j=1}^{J} \exp\left(-2n_j t_j^2\right)$. Set

$$t_j = \left(\frac{1}{2n_j} \log \frac{2J}{\alpha}\right)^{1/2} \equiv t_{j,\alpha}. \tag{5.5}$$

and

$$\mathcal{G}_\alpha = \left\{(g_1, \ldots, g_J) : \hat{g}(x_j) - p\left(x_j\right) t_{j,\alpha} \leq g_j \leq \hat{g}(x_j) + p\left(x_j\right) t_{j,\alpha} \; \forall j\right\}. \tag{5.6}$$

Then $\mathbb{P}(g_0 \in \mathcal{G}_\alpha) \geq 1 - \alpha$.

5.2.2. *Random Design.* Let X be random and Assumption 1(b) hold. Then

$$-\tau \leq (Y_{\ell,j} - \tau) I\left(X_i = x\right) \leq 1 - \tau.$$

Another application of Hoeffding's inequality gives

$$\mathbb{P}\left[\left|\hat{g}\left(x_j\right) - g_0\left(x_j\right)\right| \geq t\right] \leq 2\exp\left(-2nt^2\right)$$

for each $j = 1, \ldots, J$ and any $t > 0$. Therefore,

$$\hat{g}\left(x_j\right) - t \leq g_0\left(x_j\right) \leq \hat{g}\left(x_j\right) + t$$

for all $j = 1, \ldots, J$ and any $t > 0$ with probability at least $1 - 2J\exp\left(-2nt^2\right)$. Now set

$$t = \left(\frac{1}{2n} \log \frac{2J}{\alpha}\right)^{1/2} \equiv t_\alpha \tag{5.7}$$

and

$$\mathcal{G}_\alpha = \left\{(g_1, \ldots, g_J) : \hat{g}(x_j) - t_\alpha \leq g_j \leq \hat{g}(x_j) + t_\alpha \; \forall j\right\}. \tag{5.8}$$

Then $\mathbb{P}(g_0 \in \mathcal{G}_\alpha) \geq 1 - \alpha$.

To compare confidence regions in (5.6) and (5.8), let Assumption 1(a) hold so that $p(x_j) = n_j/n$. Then,

$$p\left(x_j\right) t_{j,\alpha} = \left[p\left(x_j\right)\right]^{1/2} t_\alpha < t_\alpha,$$

which implies that $[\hat{c}_L, \hat{c}_U]$ is a longer interval under the random design than under the fixed design. This happens because of the need to account for randomness of X when we apply Hoeffding's inequality to the random design.

5.3. **Asymptotic Inference.** This section gives asymptotic confidence regions $\mathcal{G}_\alpha$ for fixed and random designs. To minimize computational complexity (see Section 6) we focus on (hyper) rectangular regions based on the Bonferroni inequality. Ellipsoidal

regions give narrower bounds $[\hat{c}_L, \hat{c}_U]$ under certain conditions but make computation much more difficult.

5.3.1. *Fixed Design.* As in Section 5.2.1, let X be fixed and Assumption 1(a) hold. It follows from the Lindeberg-Lévy central limit theorem that for each $j = 1, \ldots, J$,

$$\left(n/n_j^{1/2}\right)(\hat{g}_j - g_{0,j}) \to_d N\left(0, \sigma_j^2\right),$$

where

$$\sigma_j^2 \equiv \mathbb{P}(Y = 1 \mid X = x_j)\left[1 - \mathbb{P}(Y = 1 \mid X = x_j)\right].$$

Estimate σ_j^2 by

$$\hat{\sigma}_j^2 \equiv \left\{n_j^{-1}\sum_{\ell=1}^{n_j} Y_{\ell,j}\right\}\left\{n_j^{-1}\sum_{\ell=1}^{n_j}(1 - Y_{\ell,j})\right\},$$

and let $z_{1-\alpha}$ denote the $(1-\alpha)$ quantile of the standard normal distribution. Set

(5.9)

$$\mathcal{G}_\alpha = \left\{(g_1, \ldots, g_J) : \hat{g}(x_j) - \frac{n_j^{1/2}\hat{\sigma}_j}{n} z_{1-\alpha/(2J)} \le g_j \le \hat{g}(x_j) + \frac{n_j^{1/2}\hat{\sigma}_j}{n} z_{1-\alpha/(2J)}\ \forall j\right\}.$$

Then the Bonferroni inequality gives $\mathbb{P}(g_0 \in \mathcal{G}_\alpha) \ge 1 - \alpha$ asymptotically.

5.3.2. *Random Design.* Let X be random and Assumption 1(b) hold. Then the Lindeberg-Lévy theorem gives

$$n^{1/2}(\hat{g}_j - g_{0,j}) \to_d N\left(0, s_j^2\right),$$

where

$$s_j^2 \equiv \mathrm{Var}\left[(Y_i - \tau) I(X_i = x_j)\right].$$

Estimate s_j^2 by

$$\hat{s}_j^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(Y_i - \tau) I(X_i = x_j) - \frac{1}{n}\sum_{i=1}^{n}(Y_i - \tau) I(X_i = x_j)\right]^2.$$

Set

$$\mathcal{G}_\alpha = \left\{(g_1, \ldots, g_J) : \hat{g}(x_j) - \frac{\hat{s}_j}{n^{1/2}} z_{1-\alpha/(2J)} \le g_j \le \hat{g}(x_j) + \frac{\hat{s}_j}{n^{1/2}} z_{1-\alpha/(2J)}\ \forall j\right\}. \tag{5.10}$$

Then, the Bonferroni inequality gives $\mathbb{P}(g_0 \in \mathcal{G}_\alpha) \ge 1 - \alpha$ asymptotically. $\mathcal{G}_\alpha$ is larger with a random design case than with a fixed design if and only if $n\hat{s}_j^2 > n_j\hat{\sigma}_j^2$.

5.4. **The Probability of Misclassification with an Option of Non-Classification.** In the population, an individual with covariate X_* is classified as $Y = 1$ if $c_L(X_*) > 0$ and as $Y = 0$ if $c_U(X_*) < 0$, where $c_L(X_*)$ and $c_U(X_*)$ are the optimal values of the objective functions defined in (3.4)–(3.5) with $r = X_*$. The individual is not classified if $c_L(X_*) \leq 0 \leq c_U(X_*)$. Thus, the *oracle classifier* assigns:

$$\text{Oracle}(X_*) = \begin{cases} 1 & \text{if } c_L(X_*) > 0, \\ 0 & \text{if } c_U(X_*) < 0, \\ \emptyset \text{ (non-classification)} & \text{if } c_L(X_*) \leq 0 \leq c_U(X_*). \end{cases}$$

The desirability of this classification rule was justified through a decision-theoretic framework in Section 4.

Using the estimation sample, an individual is classified as $Y = 1$ if $\hat{c}_L(X_*) > 0$ and as $Y = 0$ if $\hat{c}_U(X_*) < 0$, where $\hat{c}_L(X_*)$ and $\hat{c}_U(X_*)$ are the optimal values of the objective functions in (5.2)–(5.3) with $r = X_*$. The individual is not classified if $\hat{c}_L(X_*) \leq 0 \leq \hat{c}_U(X_*)$. Thus, the *maximum score classifier* assigns:

$$\text{MaximumScore}(X_*) = \begin{cases} 1 & \text{if } \hat{c}_L(X_*) > 0, \\ 0 & \text{if } \hat{c}_U(X_*) < 0, \\ \emptyset \text{ (non-classification)} & \text{if } \hat{c}_L(X_*) \leq 0 \leq \hat{c}_U(X_*). \end{cases}$$

We define the misclassification indicator at X_* by

$$\mathcal{M}(X_*) = \begin{cases} 1 & \text{if Oracle}(X_*) \neq \text{MaximumScore}(X_*), \\ 0 & \text{otherwise.} \end{cases}$$

To characterize the probability of misclassification, we allow X_* to be either fixed or random.

Assumption 4. *For the covariate X_*, $c_L(X_*) < c_U(X_*)$ (partial identification). In addition, one of the following holds:*

(a) *(i) X_* is fixed and may differ from the support points $\{x_j\}_{j=1}^J$ defined in Assumption 1(a).*
(ii) For some small $\epsilon > 0$, either $c_L(X_*) > \epsilon$ or $c_U(X_*) < -\epsilon$.

(b) *(i) X_* is random, possibly with a distribution different from that of X defined in Assumption 1(b), and X_* is independent of the estimation sample $\{X_1, \ldots, X_n\}$.*

(ii) For some constants $M_L, M_U > 0$ and for all sufficiently small $\epsilon > 0$,

$$\mathbb{P}_{X_*}\{0 < c_L(X_*) \leq \epsilon\} \leq M_L\epsilon, \quad \mathbb{P}_{X_*}\{-\epsilon \leq c_U(X_*) < 0\} \leq M_U\epsilon.$$

Part (a) considers a fixed design, while part (b) addresses a random design. In part (i) under each setting, we allow X_* or its distribution to differ from that of the training sample. Part (ii) imposes margin conditions to ensure separation in both designs. The margin conditions in part (ii) are natural in our setting, as they formalize the idea that classification becomes more reliable when $c_L(X_*)$ and $c_U(X_*)$ are well-separated from zero. When these conditions hold, small estimation errors are unlikely to alter the classification outcome, making it easier to mimic the oracle classification. In contrast, if the margin conditions fail (i.e., when either $c_L(X_*)$ or $c_U(X_*)$ lies close to zero in a fixed design, or with non-negligible probability in a random design), even minor sampling variability can result in misclassification. This undermines both the stability and the interpretability of the classification rule, particularly in finite samples.

Theorem 5.2 (Misclassification Probability Bounds)**.** *Suppose* (5.1) *holds. Then, under Assumption 4(a),*

$$\mathbb{P}\{\mathcal{M}(X_*) = 1\} \leq \alpha + \mathbb{P}(\hat{c}_L(X_*) - c_L(X_*) \leq -\epsilon) + \mathbb{P}(\hat{c}_U(X_*) - c_U(X_*) \geq \epsilon).$$

Under Assumption 4(b),

$$\begin{aligned}\mathbb{P}\{\mathcal{M}(X_*) = 1\} \leq \alpha + \mathbb{P}(\hat{c}_L(X_*) - c_L(X_*) \leq -\epsilon) + \mathbb{P}(\hat{c}_U(X_*) - c_U(X_*) \geq \epsilon) \\ + M_L\epsilon + M_U\epsilon.\end{aligned}$$

For the fixed design, the misclassification probability is bounded by two components: the coverage error α and the estimation errors $\mathbb{P}(\hat{c}_L(X_*) - c_L(X_*) \leq -\epsilon) + \mathbb{P}(\hat{c}_U(X_*) - c_U(X_*) \geq \epsilon)$. There is a trade-off between these two components: as α decreases, the interval $[\hat{c}_L(X_*), \hat{c}_U(X_*)]$ widens, leading to larger estimation errors. This mirrors the familiar size-power trade-off in hypothesis testing.

For the random design, an additional component $M_L\epsilon + M_U\epsilon$ appears, reflecting the difficulty of classification when $c_L(X_*)$ or $c_U(X_*)$ is close to zero.

5.5. **The Probability of Misclassification with Random Classification.** In this section, we consider the case of random classification. Specifically, random classification by the oracle occurs when $c_L(X_*) \leq 0 \leq c_U(X_*)$ for a given covariate value X_*. Let r be a binary random variable taking values in $\{0, 1\}$ with equal probability, i.e., $\mathbb{P}(r = 1) = \mathbb{P}(r = 0) = 0.5$, and let $Y = r$ when random classification occurs.

Define the oracle classification outcome by

$$M_{\mathrm{OR}} = \begin{cases} 1 & \text{if the oracle classifies } Y = 1, \\ 0 & \text{if the oracle classifies } Y = 0, \end{cases}$$

and the classification outcome under the maximum score rule by

$$M_{\mathrm{MS}} = \begin{cases} 1 & \text{if the maximum score model classifies } Y = 1, \\ 0 & \text{if the maximum score model classifies } Y = 0. \end{cases}$$

Then the corresponding classification rules are:

$$\begin{aligned} M_{\mathrm{OR}} &= I[c_L(X_*) > 0] + I[c_U(X_*) < 0] + r \cdot I[c_L(X_*) \le 0 \le c_U(X_*)], \\ M_{\mathrm{MS}} &= I[\hat{c}_L(X_*) > 0] + I[\hat{c}_U(X_*) < 0] + r \cdot I[\hat{c}_L(X_*) \le 0 \le \hat{c}_U(X_*)], \end{aligned} \tag{5.11}$$

where both oracle and maximum score classifiers are assumed to use the same random classification r. We redefine the misclassification indicator at X_* as:

$$\mathcal{M}(X_*) = \begin{cases} 1 & \text{if } M_{\mathrm{OR}} \neq M_{\mathrm{MS}}, \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 5.3 (Misclassification Probability Bounds with Random Classification)**.** *Suppose* (5.1) *holds. Then, under Assumption 4(a),*

$$\mathbb{P}\{\mathcal{M}(X_*) = 1\} \le \alpha + 0.5\mathbb{P}(\hat{c}_L(X_*) - c_L(X_*) \le -\epsilon) + 0.5\mathbb{P}(\hat{c}_U(X_*) - c_U(X_*) \ge \epsilon).$$

Under Assumption 4(b),

$$\begin{aligned} \mathbb{P}\{\mathcal{M}(X_*) = 1\} \le \alpha &+ 0.5\mathbb{P}(\hat{c}_L(X_*) - c_L(X_*) \le -\epsilon) + 0.5\mathbb{P}(\hat{c}_U(X_*) - c_U(X_*) \ge \epsilon) \\ &+ 0.5M_L\epsilon + 0.5M_U\epsilon. \end{aligned}$$

The bounds in Theorem 5.3 differ from those in Theorem 5.2 due to the role of random classification. In the setting of Theorem 5.2, abstention is permitted, so it is possible for the oracle to classify $Y = 0$ or $Y = 1$, while the maximum score classifier abstains. This leads to a classification disagreement. In contrast, under random classification (as in Theorem 5.3), abstention is not an option, and both the oracle and the maximum score classifier randomly assign $Y = 0$ or $Y = 1$ with equal probability when the decision is ambiguous. Consequently, they agree with probability 0.5 in such cases, and this probabilistic agreement is reflected in the factor of 0.5 appearing in the bounds in Theorem 5.3.

Remark 8 (Randomization Device). In (5.11), it is important that both the oracle and maximum score classifiers use the same randomization device, denoted by r. It is also possible to consider a setting where the oracle and the maximum score classifier use independent randomization devices (e.g., flipping fair coins independently). This can be described as:

$$M_{\text{OR}} = I[c_L(X_*) > 0] + I[c_U(X_*) < 0] + r_{\text{OR}} \cdot I[c_L(X_*) \leq 0 \leq c_U(X_*)],$$
$$M_{\text{MS}} = I[\hat{c}_L(X_*) > 0] + I[\hat{c}_U(X_*) < 0] + r_{\text{MS}} \cdot I[\hat{c}_L(X_*) \leq 0 \leq \hat{c}_U(X_*)],$$

where r_{OR} and r_{MS} are independent binary random variables with probability 0.5. Under this setup, the conclusions of Theorem 5.3 are modified. For the fixed design:

$$\mathbb{P}\{\mathcal{M}(X_*) = 1\} \leq \alpha + 0.5\mathbb{P}(\hat{c}_L(X_*) - c_L(X_*) \leq -\epsilon) + 0.5\mathbb{P}(\hat{c}_U(X_*) - c_U(X_*) \geq \epsilon) + 0.5I[c_L(X_*) \leq 0 \leq c_U(X_*)],$$

and for the random design:

$$\mathbb{P}\{\mathcal{M}(X_*) = 1\} \leq \alpha + 0.5\mathbb{P}(\hat{c}_L(X_*) - c_L(X_*) \leq -\epsilon) + 0.5\mathbb{P}(\hat{c}_U(X_*) - c_U(X_*) \geq \epsilon) + 0.5M_L\epsilon + 0.5M_U\epsilon + 0.5\mathbb{P}[c_L(X_*) \leq 0 \leq c_U(X_*)].$$

These bounds can be derived using arguments similar to those in the proof of Theorem 5.3, with appropriate adjustments to the analysis of Case 3.

6. Computational Algorithms

In this section, we describe how to solve the optimization problems given in (5.2)-(5.3). In all the cases we considered, $\mathcal{G}_\alpha$ has the form

$$\mathcal{G}_\alpha = \{(g_1, \ldots, g_J) : \hat{g}(x_j) - \hat{s}(x_j, \alpha) \leq g_j \leq \hat{g}(x_j) + \hat{s}(x_j, \alpha)\ \forall j\} \tag{6.1}$$

for an appropriate choice of $\hat{s}(x_j, \alpha) > 0$. For example, for asymptotic inference under the random design, we have $\hat{s}(x_j, \alpha) = n^{-1/2}\hat{s}_j z_{1-\alpha/(2J)}$ (see the form of $\mathcal{G}_\alpha$ defined in (5.10)). The constraints in (5.3) are bilinear and could potentially create computational difficulties; however, it turns out that it is easy to solve (5.2)-(5.3) with $\mathcal{G}_\alpha$ in the form of (6.1). Specifically, the following theorem shows that it suffices to solve a linear programming (LP) problem, which is known to be highly scalable.

Theorem 6.1. *Solutions to* (5.2)-(5.3) *with* $\mathcal{G}_\alpha$ *in the form of* (6.1) *can be obtained by solving*

$$\underset{b \in \mathbf{B}}{\text{maximize}} \left(\underset{b \in \mathbf{B}}{\text{minimize}} \right) : \ r'b \tag{6.2}$$

subject to

$$(x_{j,1} + b_2 x_{j,2} + \cdots + b_q x_{j,q}) \geq 0 \quad \textit{if } \hat{g}(x_j) - \hat{s}(x_j, \alpha) > 0, \tag{6.3a}$$

$$(x_{j,1} + b_2 x_{j,2} + \cdots + b_q x_{j,q}) \leq 0 \quad \textit{if } \hat{g}(x_j) + \hat{s}(x_j, \alpha) < 0. \tag{6.3b}$$

This result highlights a key computational simplification: for each j, only one of the two linear constraints in (6.3) needs to be imposed, and only when $\hat{g}(x_j)$ deviates sufficiently from zero. If $\hat{g}(x_j)$ lies within the interval $[-\hat{s}(x_j, \alpha), \hat{s}(x_j, \alpha)]$, no constraint is imposed for that index. This selective enforcement leads to an LP problem with fewer active constraints, while still exactly solving the original bilinear formulation.

7. Continuous Covariates

In this section, we extend our framework to Manski and Tamer (2002, MT hereafter) type bounds for maximum score estimation when a continuously distributed vector is only observed or treated as only observed to be in a hypercube.

7.1. **Model.** Let X be a discretely distributed random vector with $\dim(X) = q_X$ and V be a continuously distributed random vector with $\dim(V) = q_V$. The model is given by

$$Y = \begin{cases} 1 & \text{if } \beta' X + \delta' V - U \geq 0 \\ 0 & \text{if } \beta' X + \delta' V - U < 0 \end{cases},$$

where the unobserved random variable U satisfies $\mathbb{P}(U \leq 0 \mid X, V) = \tau$ almost surely for some constant $0 < \tau < 1$. We use the scale/sign normalization $\delta_1 = 1$, where δ_1 is the first element of δ. A key assumption is that V is unknown (or treated as unknown and we discretize V in advance) but is in the hypercube $V_0 \leq V \leq V_1$ component-wise, where V_0 and V_1 are known, non-stochastic constants or observed discrete random variables. In this setting, $\theta \equiv (\beta', \delta')'$ is not point identified. The objective here is to estimate upper and lower bounds on the identified interval for $r'\theta$, where r is a constant vector whose components are not all zero.

Let $W \equiv (X', V_0', V_1')'$ denote the observed covariates. We make the following regularity conditions.

Assumption 5. (1) *W is discretely distributed.*

(2) *$\{(Y_i, W_i) : i = 1, \ldots, n\}$ is an independent random sample of (Y, W).*

(3) *$\theta \equiv (\beta', \delta')' \in \Theta \subset \mathbb{R}^{q_X + q_V}$, where Θ is a known compact set with the restriction that $\delta_1 = 1$.*

We now make the following generalization of MT's interval, monotonicity, and mean independence (IMMI) assumptions.

Assumption 6 (IMMI)**.** (1) *Interval (I):* $V_0 \leq V \leq V_1$ *if* V_0 *and* V_1 *are non-stochastic, and* $\mathbb{P}(V_0 \leq V \leq V_1) = 1$ *if they are random, where the inequalities hold component-wise.*

(2) *Monotonicity (M):* $\mathbb{E}(Y \mid X, V)$ *exists and is weakly increasing in each component of* V*. Equivalently, the components of* δ *are non-negative.*

(3) *Mean independence (MI):* $\mathbb{E}(Y \mid X, V, V_0, V_1) = \mathbb{E}(Y \mid X, V)$.

The monotonicity assumption requires the signs of the coefficients of the components of V to be known, which is often the case in applications (e.g., if V is a price). Given this knowledge, there is no further loss of generality in assuming that the signs are non-negative; a negative sign can be turned into a positive one by changing the sign of the relevant component of V.

7.2. **Identification Result.** Let $\mathcal{W} \equiv \{w_j \equiv (x_j, v_{0j}, v_{1j}) : j = 1, \ldots, J\}$ denote mass points of W. The following theorem simplifies and generalizes Proposition 1 of MT.

Lemma 7.1. *Let Assumption 6 (that is, the IMMI assumptions) hold. Then, the following sharp bounds hold for all* $w \equiv (x, v_0, v_1) \in \mathcal{W}$*:*

$$\mathbb{E}(Y \mid X = x, V = v_0) \leq \mathbb{E}(Y \mid W = w) \leq \mathbb{E}(Y \mid X = x, V = v_1). \tag{7.1}$$

Now define

$$f(w_j) \equiv \mathbb{P}(Y = 1 \mid X = x_j, V_0 = v_{0j}, V_1 = v_{1j}) - \tau$$

and

$$d_j = \operatorname{sgn}[f(w_j)] = \begin{cases} 1 & \text{if } f(w_j) > 0 \\ 0 & \text{if } f(w_j) = 0 \\ -1 & \text{if } f(w_j) < 0 \end{cases}.$$

We now make an analog of Assumption 2.

Assumption 7. *There is an* $\varepsilon > 0$ *such that*

$$\min_{1 \leq j \leq J} \left(|\beta' x_j + \delta' v_{0j}|, |\beta' x_j + \delta' v_{1j}| \right) = \varepsilon$$

for all $j = 1, \ldots, J$.

Under Assumption 7, we have that $|f(w_j)| > 0$ for all $j = 1, \ldots, J$. This is necessary to exclude the case of $f(w_j) = 0$ as we have explained in Remark 3. Note that by Lemma 7.1 and the fact that $[\mathbb{P}(Y = 1 \mid X = x_j, V = v_j) - \tau] > 0 \Leftrightarrow \beta' x_j + \delta' v_j > 0$,

$$\begin{aligned} d_j = 1 &\Longrightarrow [\mathbb{P}(Y = 1 \mid X = x_j, V_0 = v_{0j}, V_1 = v_{1j}) - \tau] > 0 \\ &\Longrightarrow [\mathbb{P}(Y = 1 \mid X = x_j, V = v_{1j}) - \tau] > 0 \\ &\Longrightarrow \beta' x_j + \delta' v_{1j} \geq \varepsilon \end{aligned}$$

and

$$\begin{aligned} d_j = -1 &\Longrightarrow [\mathbb{P}(Y = 1 \mid X = x_j, V_0 = v_{0j}, V_1 = v_{1j}) - \tau] < 0 \\ &\Longrightarrow [\mathbb{P}(Y = 1 \mid X = x_j, V = v_{0j}) - \tau] < 0 \\ &\Longrightarrow \beta' x_j + \delta' v_{0j} \leq -\varepsilon. \end{aligned}$$

Therefore, the identified set for θ is

$$\left\{\theta \equiv (\beta', \delta')' \in \Theta : I(d_j = 1)(\beta' x_j + \delta' v_{1j}) - I(d_j = -1)(\beta' x_j + \delta' v_{0j}) \geq \varepsilon \; \forall j = 1, \ldots, J\right\}.$$

The identified upper and lower bounds on $r'\theta$ are optimal values of the objective functions of the linear program

$$\underset{\theta \in \Theta}{\text{maximize}} \left(\underset{\theta \in \Theta}{\text{minimize}} \right) : \; r'\theta \tag{7.2}$$

subject to

$$I(d_j = 1)(\beta' x_j + \delta' v_{1j}) - I(d_j = -1)(\beta' x_j + \delta' v_{0j}) \geq \varepsilon \tag{7.3}$$

for all $j = 1, \ldots, J$. As in Remark 2, outer bounds can be obtained by replacing ε with zero.

7.3. **Inference.** For brevity, we focus on the random design case. It is straightforward to modify the following discussion to the fixed design case. As before, we define

$$g(w_j) \equiv [\mathbb{P}(Y = 1 \mid X = x_j, V_0 = v_{0j}, V_1 = v_{1j}) - \tau] \, p_W(w_j),$$

where $p_W(\cdot)$ is the probability mass function of W. As $p_W(w_j) > 0$, we have that $\text{sgn}[f(w_j)] = \text{sgn}[g(w_j)]$ for all $j = 1, \ldots, J$. We now modify the estimator by

$$\hat{g}(w_j) = n^{-1} \sum_{i=1}^{n} (Y_i - \tau) \, I(W_i = w_j)$$

and assume that we have the following confidence set for $g_j \equiv g(w_j)$:

$$\mathcal{G}_\alpha = \{(g_1, \ldots, g_J) : \hat{g}(w_j) - \hat{s}(w_j, \alpha) \leq g_j \leq \hat{g}(w_j) + \hat{s}(w_j, \alpha)\ \forall j\} \tag{7.4}$$

for an appropriate choice of $\hat{s}(w_j, \alpha) > 0$.

Following the arguments used to prove Theorem 6.1, we can show that it is possible to conduct inference by solving

$$\underset{\theta \in \Theta}{\text{maximize}} \left(\underset{\theta \in \Theta}{\text{minimize}} \right) : \ r'\theta \tag{7.5}$$

subject to

$$(\beta' x_j + \delta' v_{1j}) \geq 0 \ \text{ if } \hat{g}(w_j) - \hat{s}(w_j, \alpha) > 0, \tag{7.6a}$$

$$(\beta' x_j + \delta' v_{0j}) \leq 0 \ \text{ if } \hat{g}(w_j) + \hat{s}(w_j, \alpha) < 0. \tag{7.6b}$$

When $v_{0j} = v_{1j}$, the resulting optimization problem is the same as (6.2).

7.4. **Classification.** Suppose that we want to classify a new member of the population using their covariate vector, say $W_* \equiv (X_*, V_*)$, where W_* may not be an element in the support $\mathcal{W}$ of W. In particular, V_* can be different from the realized values of (V_{0i}, V_{1i}). Then, we solve (7.5) with $r = W_*$. As before, we propose to assign $(Y = 0)$ if the solution to the minimization problem is negative; $(Y = 1)$ if the solution to the maximization problem is positive; and abstain from classification or apply random classification otherwise.

8. Empirical Illustrations

In Section 8.1, we apply our method to classify resumes in the context of hiring decisions. In Section 8.2, we demonstrate how the approach can be extended to interval-valued covariates using the resume classification example. Finally, in Section 8.3, we use our method to assess the likelihood of marriage among women in developing countries.

8.1. **Incentivized Resume Rating.** In this section, we use data from Kessler, Low, and Sullivan (2019, KLS hereafter) to illustrate the methodology developed in this paper. KLS introduced a new experimental paradigm, called incentivized resume rating (IRR), which invites employers to evaluate resumes known to be hypothetical (avoiding deception) and provides incentives by matching employers with real job seekers. For each resume, KLS asked two questions, both on a Likert scale from 1 to 10:

(i) "How interested would you be in hiring [Name]?" (1 = "Not interested"; 10 = "Very interested");

(ii) "How likely do you think [Name] would be to accept a job with your organization?" (1 = "Not likely"; 10 = "Very likely").

We construct the binary response variable by setting $Y_i = 1$ if the rating for the first question is higher than or equal to 5 and 0 otherwise. Among possible covariates, we include:

- GPA $\in \{3.0, 3.2, \ldots, 4.0\}$, that is, GPA is grouped into six categories via `ceiling`$(\text{GPA} \times 5)/5$;
- An indicator variable for Top Internship, which refers to internships at prestigious firms;
- Intercept.

GPA and Top Internship are two most important predictors in the KLS regression analysis (see Table 1 in KLS). In this example, $n = 2880$ and $J = 6 \times 2 = 12$. Out of 12 covariate groups, the minimum, median, and maximum group sizes are 110, 226, and 347, respectively.

Figure 1 displays the sample probabilities $\{\mathbb{P}(Y = 1|X = x_j) : j = 1, \ldots, J\}$ of receiving the rating of 5 or higher for hiring interest by covariate groups. Blue triangles correspond to 6 GPA groups (points of support of the covariates) with a top internship (i.e., Top Internship = 1) and red squares those without a top internship (i.e., Top Internship = 0). It can be seen that having a top internship boosts the ratings although the vertical distance between the triangle and square is not uniform across different GPA groups. The orange horizontal line represents the probability of 0.5, suggesting that it would be interesting to consider $\tau = 0.5$ in this example.

In view of Figure 1, we consider the following simple linear specification for $\beta' x$ with $\tau = 0.5$:

$$\beta' x = \beta_1 \times GPA + \beta_2 \times Top\ Internship + \beta_3, \tag{8.1}$$

where $\beta_1 \equiv 1$ by scale normalization. The main parameter of interest is β_2, which measures the effectiveness of having a top internship relative to the effect of GPA.

TABLE 1. Interval estimates for coefficients across different methods

	(1) Logit	(2) MS: Sample	(3) MS: Fixed Design	(4) MS: Random Design
Top Internship	[0.32, 0.58]	[0.2, 0.6]	[0.0, 1.0]	[−0.2, 1.0]
Intercept	[−3.73, −3.59]	[−3.6, −3.4]	[−4.0, −3.4]	[−4.0, −3.4]

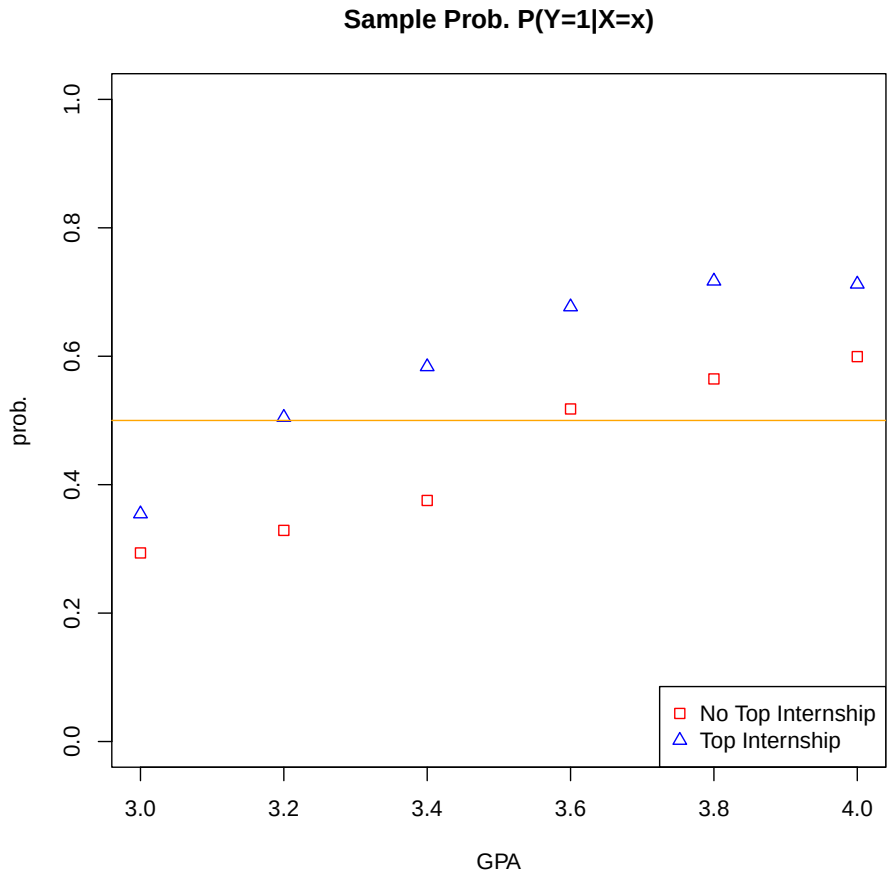

FIGURE 1. $\mathbb{P}(Y = 1|X = x)$ by GPA and Top Internship

Table 1 presents interval estimates for β_2 and β_3 across different methods. In column (1), a logit model based on (8.1) is fitted and the misspecification-robust asymptotic 95% confidence intervals for β_j/β_1 for $j = 2, 3$ are given. In columns (2)-(4), maximum score estimation of (8.1) is considered with $\tau = 0.5$ under the restriction that $-10 \leq \beta_j \leq 10$ for $j = 2, 3$. In column (2), (6.2)-(6.3) is solved for $r'\beta = \beta_j$ with $\hat{s}(x_j, \alpha) \equiv 0$; that is, we treat the $\hat{g}_j$'s as if they were true probabilities. In columns (3) and (4), we use the asymptotic 95% confidence regions for the fixed and random designs, respectively. The sample size was too small to consider finite sample inference in this example. In Appendix B, we present results from small Monte Carlo experiments comparing finite and asymptotic inference methods. In summary, a substantially larger sample size is required to obtain informative bounds. It can be seen from Table 1 that the effect of Top Internship is strictly positive in columns (1)

and (2) and has relatively tight intervals. The intervals are wider in columns (3) and (4).

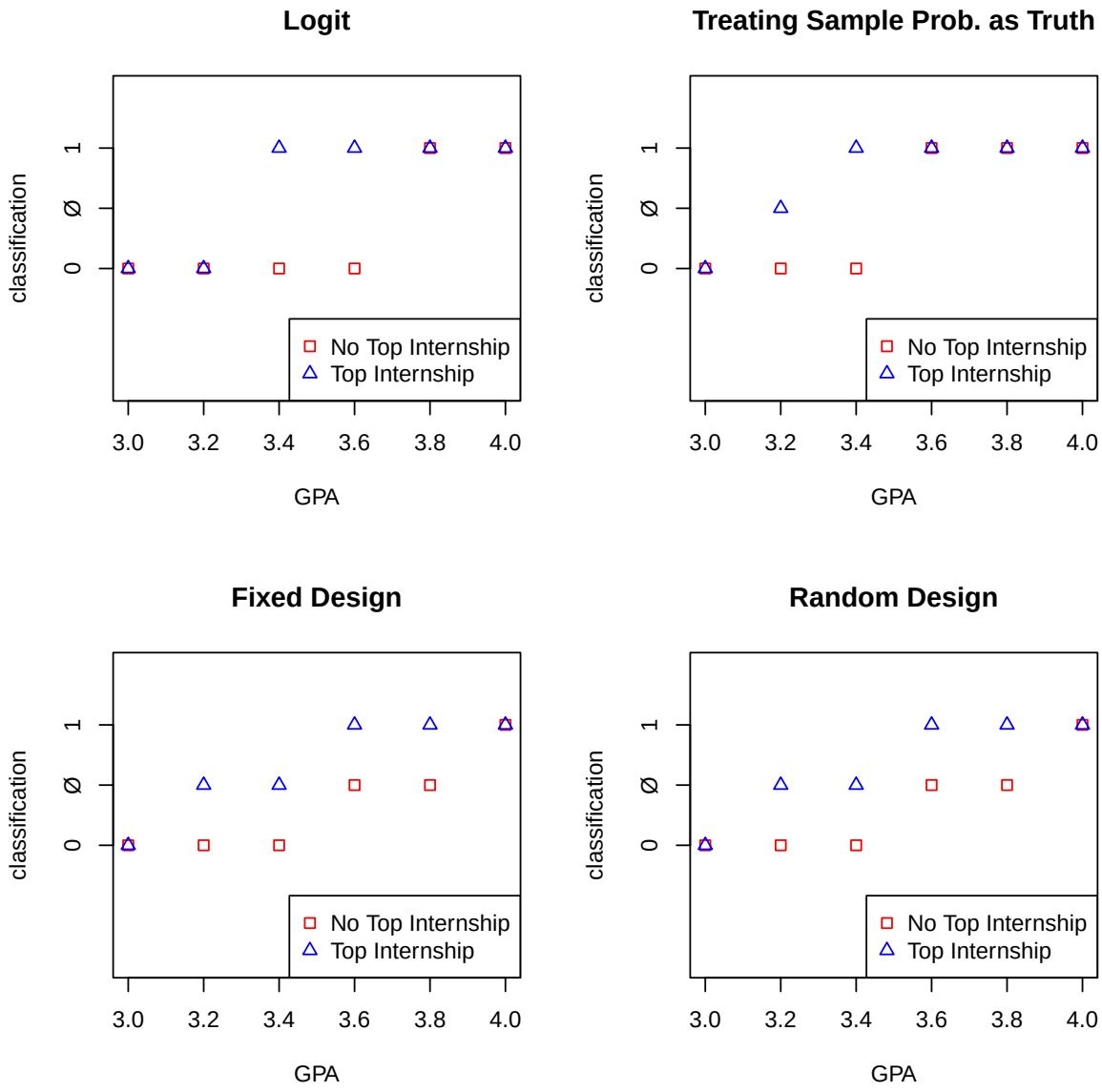

FIGURE 2. Classification by Covariate Groups

We now move to our main focus, namely, the classification exercise for each covariate group. Figure 2 shows the results of classification by covariate groups. In the top left panel, we classify each covariate group by $I(x_j'\hat{\beta}_{\text{Logit}} > \tau)$, where $\hat{\beta}_{\text{Logit}}$ is the logit estimate. That is, if the predicted probability is greater than $\tau = 0.5$, we assign 1 and

if it is less than $\tau = 0.5$, we assign 0. In other panels, we use maximum score methods and classify the covariate groups as described in Section 5.4. Here, $\emptyset$ refers to the case of "non-classification." The maximum score methods used in Figure 2 match those in columns (2)-(4) in Table 1 with a suitable choice of $r'\beta$. The classification results using the logit model suggest that Top Internship is a game changer for the middle GPA groups of 3.4 and 3.6. The top right panel that treats sample probabilities as the truth show a similar patten, although it indicates that Top Internship is crucial for GPA groups of 3.2 and 3.4. The bottom panels are identical, implying that there is no difference in classification between the fixed and random design cases. The results are more prudent here as there are 4 incidences of non-classification. The GPA groups of 3.6 or higher with a top internship are classified as 1, whereas the GPA groups of 3.4 or lower without a top internship are classified as 0. Finally, we do not display classification outcomes for the abstention region $\emptyset$ under random classification, as its value will be either 0 or 1 depending on the outcome of the randomization device.

8.2. **Incentivized Resume Rating: GPA as a Interval Covariate.** In this subsection, we illustrate the extension discussed in Section 7. Recall that we have grouped GPA into six categories via `ceiling`$(\text{GPA} \times 5)/5$. In the original data, the range of GPA is $[2.90, 4.00]$ and it is recorded in two decimal points. Thus, we now view GPA as an interval covariate and it takes the following interval values: $\{(v_{0j}, v_{1j})_{j=1}^{6}\} = \{[2.9, 3.0], [3.0, 3.2], \ldots, , [3.8, 4.0]\}$.

TABLE 2. Interval estimates for coefficients: GPA as Truth vs. as Interval Data

	(1)	(2)	(3)	(4)
	GPA as Truth		GPA as Interval Data	
	Fixed Design	Random Design	Fixed Design	Random Design
Top Internship	[0.0, 1.0]	[−0.2, 1.0]	[−0.2, 1.1]	[−0.4, 1.1]
Intercept	[−4.0, −3.4]	[−4.0, −3.4]	[−4.0, −3.2]	[−4.0, −3.2]

Table 2 reports of the effects of treating GPA as an interval covariate. Columns (1) and (2) reproduce the estimates in Table 2, where we treat GPA as "truly" discretely distributed; columns (3) and (4) provide new estimates. It can be seen that the resulting bounds on Top Internship become larger to reflect the nature of interval censoring.

We also revisit our classification exercise. Figure 3 compares classification results between the baseline results in Figure 2 and the new results treating GPA as an interval covariate. The top panel sub-figures reproduce the estimates in Figure 2 and

the bottom panel sub-figures show equivalent classification results, only changing GPA to be an interval covariate. Unsurprisingly, there are more cases of non-classification ($\emptyset$ in the y-axis) as the bounds are wider with interval censoring. As before, we do not display classification outcomes under random classification.

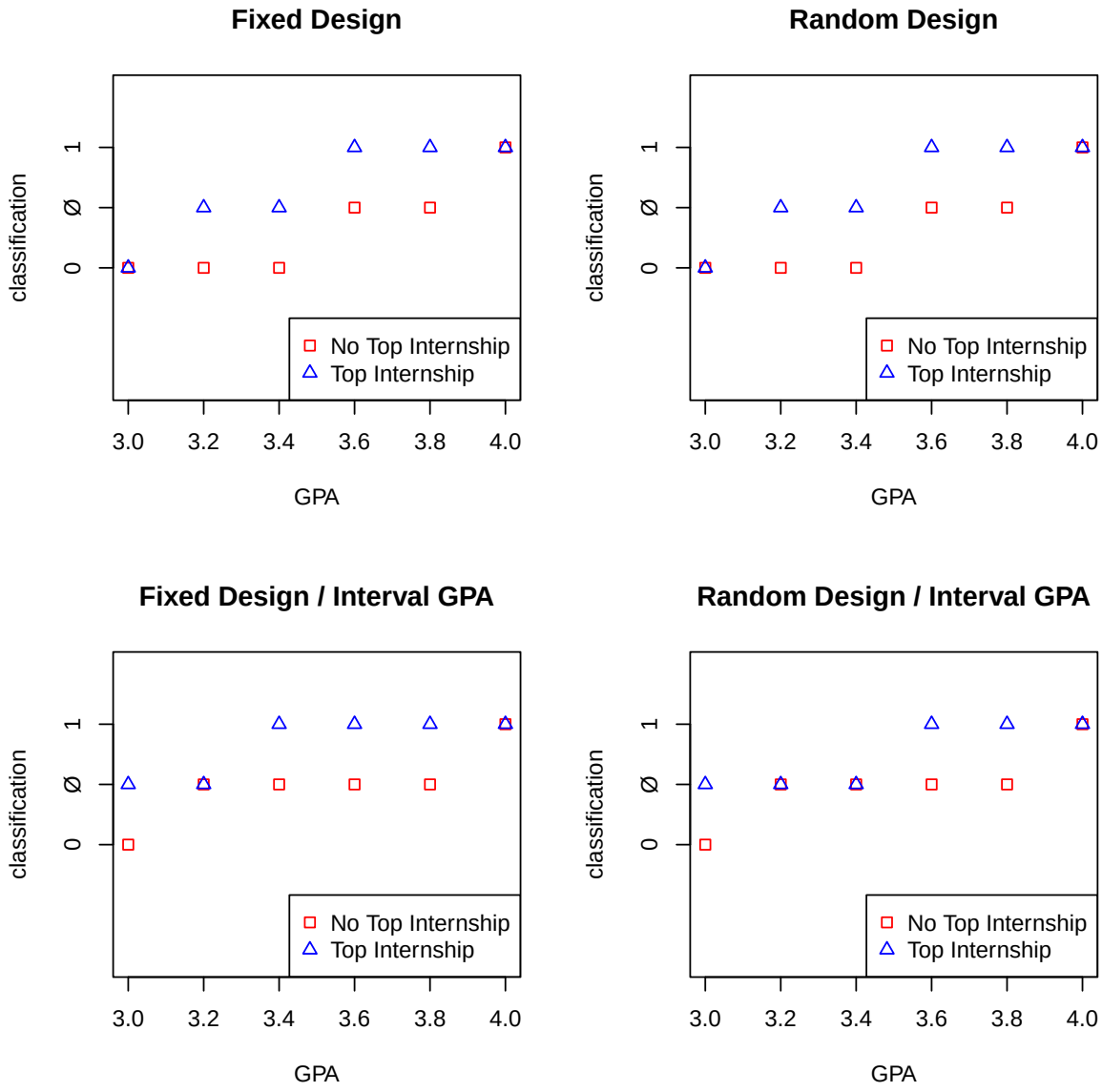

FIGURE 3. Classification by Covariate Groups: GPA as Truth vs. as Interval Data

8.3. **Age of Marriage.** In this subsection, we provide a second empirical example using data from Corno, Hildebrandt, and Voena (2020). In particular, we use their Sub-Saharan Africa (SSA) sample, which contains more than 2 million observations with survey sampling weights. Furthermore, it is unlikely that observations are independent across each other. In the regression analysis of Corno, Hildebrandt, and Voena (2020), regression coefficients are estimated with survey sampling weights and the standard errors are clustered at the grid cell level. Our framework can be extended to account for clustered dependence and survey sampling weights. Since these details are not essential for understanding the empirical results in this section, we provide them in Appendix C. In particular, we focus on asymptotic inference in Section C.2, as the effective sample size was too small to apply the finite sample inference method.

In this example, the dependent variable is a binary variable for marriage, coded to 1 if a woman married at the age corresponding to the observation and 0 otherwise. Among possible covariates, we include:

- age in years (minimum: 12 and maximum: 24);
- an indicator variable for drought, which is the main covariate in Corno, Hildebrandt, and Voena (2020);
- 3 birth decade fixed effects: the omitted group is the oldest;
- an intercept term.

They are all discrete covariates and the resulting J is $J = 13 \times 2 \times 4 = 104$. Figure 4 displays the sample probabilities $\{\mathbb{P}(Y = 1|X = x_j) : j = 1, \ldots, J\}$ of marriage by covariates: age, drought, and birth cohorts. The sample probabilities are computed using country population-adjusted survey sampling weights. The orange horizontal line in each panel represents the probability of 0.12, which is the value of τ in this example.

It can be seen from Figure 4 that the age effects tend to be linear up to around age 18 and become more or less flat after age 18. In view of this, we consider the following specification for $\beta' x$ with $\tau = 0.12$:

$$\begin{aligned} \beta' x = {} & \beta_1 \times (age - 18) I(age \leq 18) + \beta_2 \times drought + \beta_3 \times I(cohort = 1960) \\ & + \beta_4 \times I(cohort = 1970) + \beta_5 \times I(cohort = 1980) + \beta_6, \end{aligned} \tag{8.2}$$

where $\beta_1 \equiv 1$.

In Table 3, we report interval estimates using our methods. The sample maximum score interval estimates are quite tight in column (1). By the discrete nature of covariates (in particular, all the regressors can take only integer values), the tightest

FIGURE 4. Probability of marriage by covariates

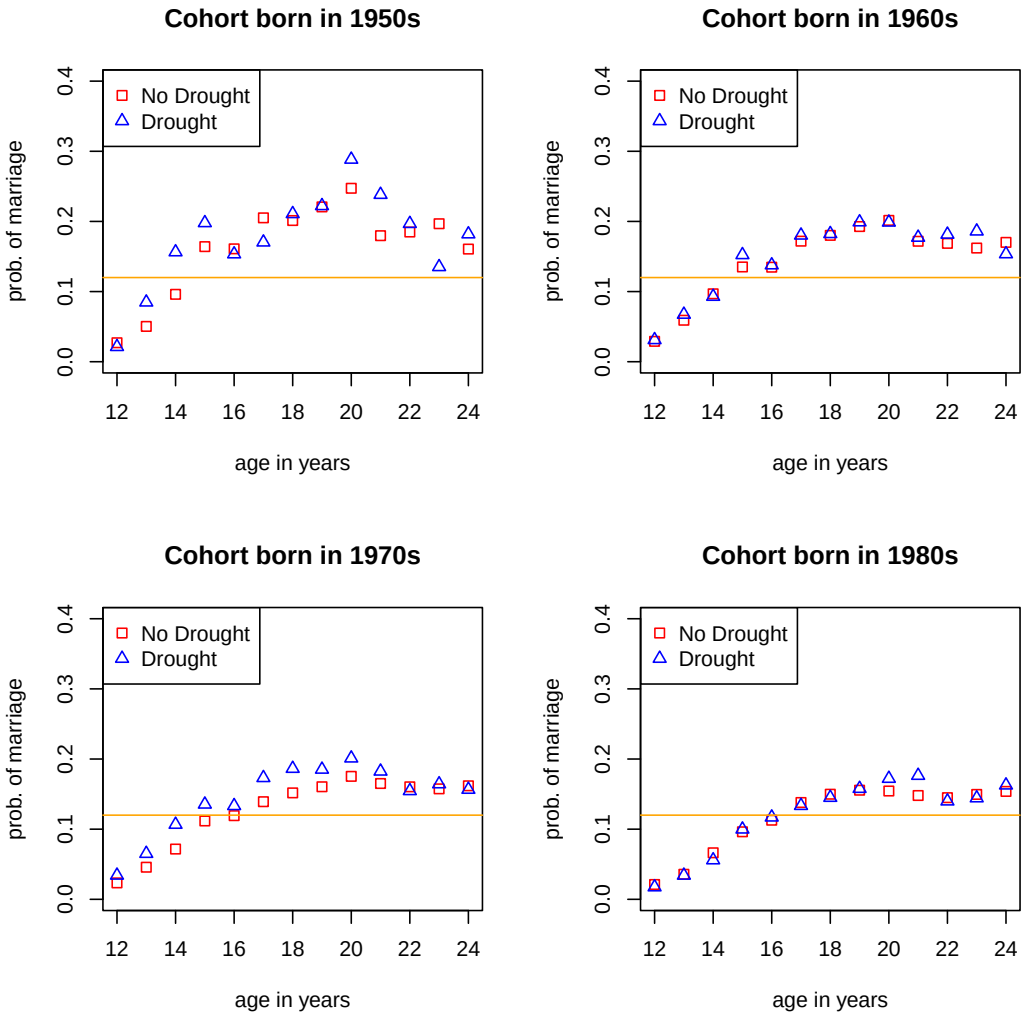

TABLE 3. Interval estimates for coefficients

	(1) MS: Sample	(2) MS: Confidence Region
Drought	[1, 1]	[−1, 3]
Birth Cohort 60	[−1, 0]	[−3, 1]
Birth Cohort 70	[−2, −1]	[−4, 1]
Birth Cohort 80	[−3, −2]	[−4, 0]
Intercept	[3, 4]	[3, 4]

interval will be a point as in the case of drought and the second tightest interval will be an interval of length 1, which is the case for all the other covariates. The maximum score interval estimates become wider in column (2) as we take into consideration sampling uncertainty by taking $\mathcal{G}_\alpha$ to be a 95% confidence region (note that $\mathcal{G}_\alpha$ is identical between the fixed and random designs here; see Appendix C.2 for details).

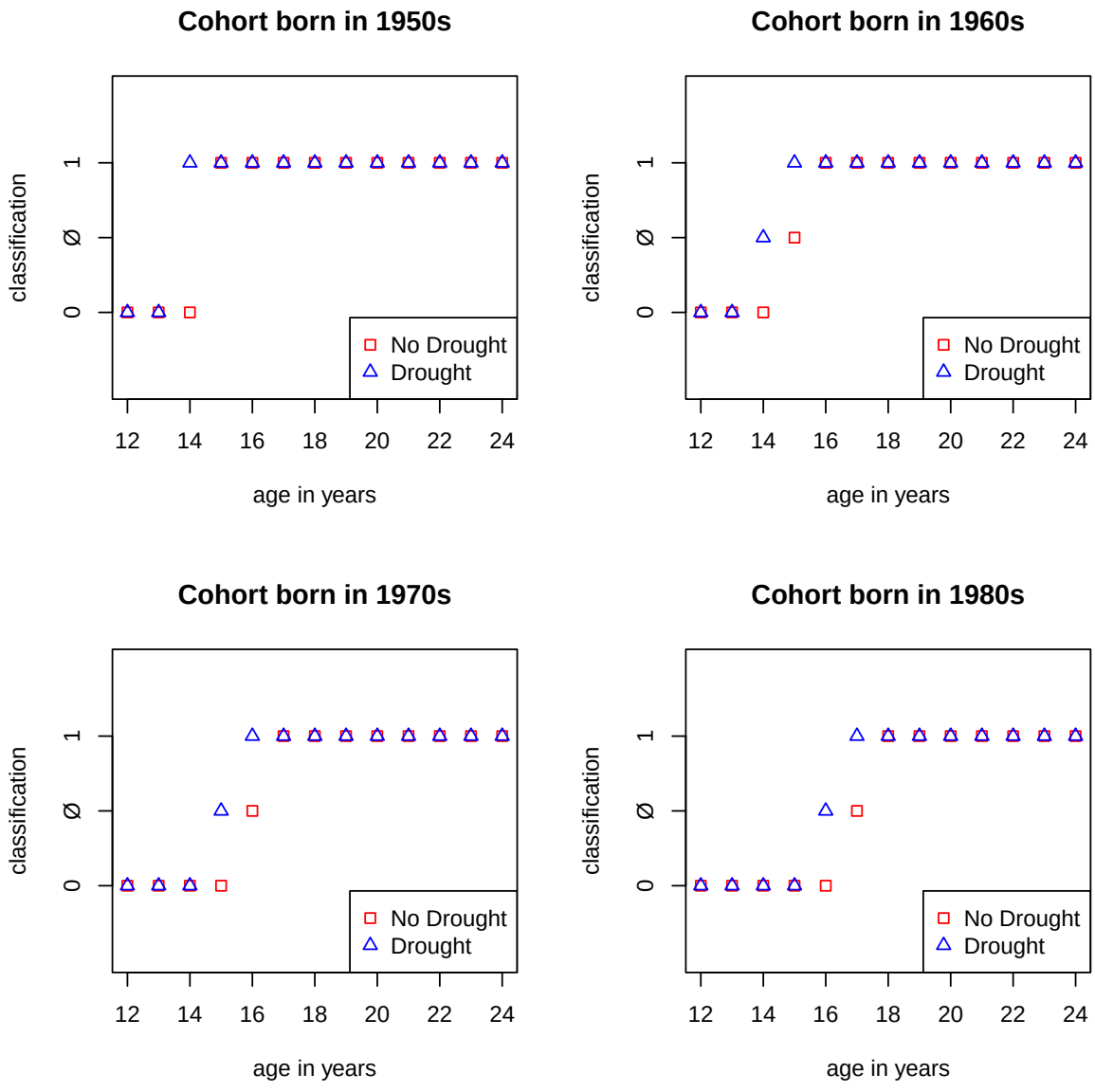

FIGURE 5. Maximum Score Classification: Treating Sample Prob. as Truth

We now move to the classification exercise. Figure 5 shows the maximum score classification results treating sample probabilities as population probabilities. One striking patten is the existence of the cohort effects, namely, "1" classifications kicked in the later ages as women were born later. The effect of drought is rather limited in the sense that it makes differences only for earlier ages. For example, the incidence of drought moves up the classification rating at age 14 for cohort 1950; ages 14 and 15 for cohort 1960; ages 15 and 16 for cohort 1970; and ages 16 and 17 for cohort 1980. Figure 6 now depicts the maximum score classification results using the 95% confidence region. There are more occurrences of non-classification but the effects of drought on classification are visible in 5 instances. Overall, the classification exercises demonstrate the usefulness of our methodology. As before, we do not display classification outcomes under random classification.

9. Monte Carlo Experiments for Classification Errors

To evaluate the effectiveness of our proposed classification rules, we conduct a Monte Carlo experiment. Specifically, we base our experimental design on the first empirical example reported in Section 8.1. We retain the same covariate design (X) from Section 8.1 and generate Y as follows:

$$\begin{aligned} Y_i &= I(Y_i^* \geq 0), \\ Y_i^* &= \beta_1 \times GPA_i + \beta_2 \times \text{Top Internship}_i + \beta_3 + U_i, \end{aligned}$$

where $U_i = \sigma(X_i)V_i$, with $V_i \sim N(0,1)$, and $(\beta_0^{(1)}, \beta_0^{(2)}, \beta_0^{(3)}) = (1, 0.4, -3.7)'$. We consider both a homoskedastic design ($\sigma(X_i) = 0.5$) and a heteroskedastic design ($\sigma(X_i) = 0.2[1+(GPA_i/3+\text{Top Internship}_i)^2]$). The parameter values of $(\beta_0^{(1)}, \beta_0^{(2)}, \beta_0^{(3)})$ are chosen to mimic the empirical results reported earlier and to ensure that the population values of $g_0(x_j)$ are never zero. The quantile of interest is $\tau = 0.5$, the confidence level is $\alpha = 0.05$, and the number of Monte Carlo repetitions is 1,000.

Table 4 reports the Monte Carlo frequencies of misclassification. Panel A considers the setting in which abstention is permitted. At the population level, an individual with covariate X_* is classified as $Y = 1$ if $c_L > 0$, as $Y = 0$ if $c_U < 0$, and remains unclassified if $c_L \leq 0 \leq c_U$. In the sample, classification follows the same rule, using the estimates $\hat{c}_L$ and $\hat{c}_U$, which are constructed based on the fixed-design asymptotic critical value. A classification error occurs whenever the sample classification differs from the population classification. This setup corresponds to the theoretical discussion of misclassification probabilities in Section 5.4.

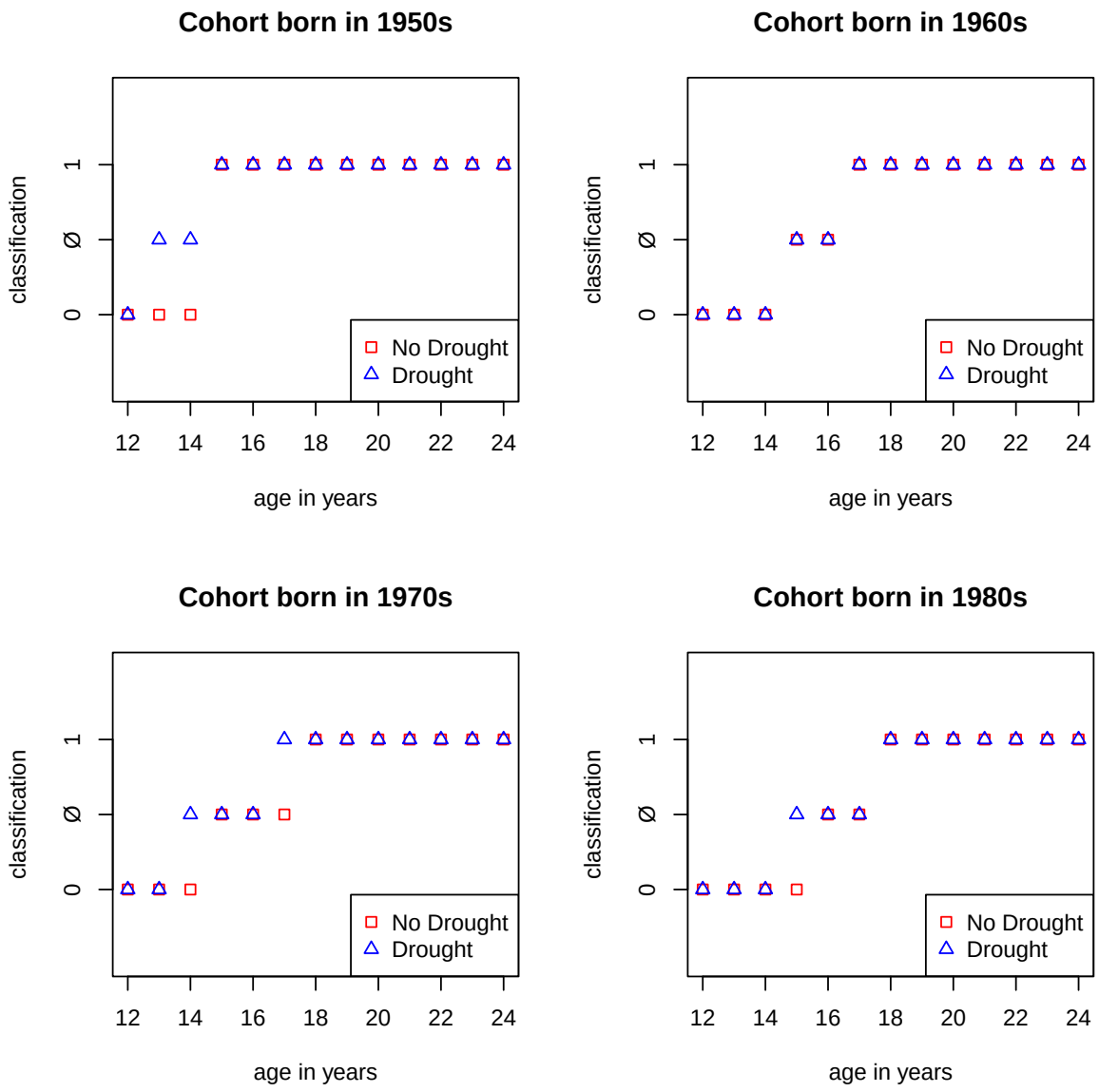

FIGURE 6. Maximum Score Classification: 95% Confidence Region

In Panel A, each row in the table corresponds to a different sample size, where increasing the sample size by a factor of c corresponds to adding $c-1$ additional copies of X, with $c = 1, 2, 3$. The column "Avg. (Pointwise)" reports the average misclassification error across $J = 12$ support points (six GPA values for each Top Internship status). Specifically, in each Monte Carlo repetition and for each support

TABLE 4. Monte Carlo Simulation Results: Classification Errors

Panel A. Abstention Allowed				
Sample Size	Homoskedastic		Heteroskedastic	
	Avg. (Pointwise)	All (Uniform)	Avg. (Pointwise)	All (Uniform)
$n = 2880$	0.058	0.608	0.157	0.901
$2n = 5760$	0.017	0.194	0.041	0.415
$3n = 8640$	0.003	0.035	0.009	0.101

Panel B. Random Classification				
Sample Size	Homoskedastic		Heteroskedastic	
	Avg. (Pointwise)	All (Uniform)	Avg. (Pointwise)	All (Uniform)
$n = 2880$	0.027	0.310	0.081	0.653
$2n = 5760$	0.008	0.096	0.021	0.235
$3n = 8640$	0.001	0.018	0.004	0.048

Panel C. Sample Frequency Classification				
Sample Size	Homoskedastic		Heteroskedastic	
	Avg. (Pointwise)	All (Uniform)	Avg. (Pointwise)	All (Uniform)
$n = 2880$	0.127	0.891	0.130	0.885
$2n = 5760$	0.125	0.880	0.125	0.885
$3n = 8640$	0.123	0.868	0.125	0.086

Notes: Here, $n = 2880$ refers to the original sample size in the KLS example.

point, we check whether the sample classification differs from the population classification, and then average these discrepancies across both repetitions and support points. As an alternative measure, the column "All (Uniform)" reports whether, for each Monte Carlo repetition, any misclassification occurs across all support points. Since "All (Uniform)" imposes a stricter criterion, the reported misclassification errors are higher in this column.

Although the exact misclassification error levels depend on the noise structure in U_i, the results show a clear trend: both measures of misclassification error decrease as the sample size increases. This illustrates that as the sample size grows, population bounds increasingly dominate sampling errors, reinforcing the reliability of the proposed classification approach.

Panel B in Table 4 considers the scenario in which abstention is not allowed but random classification is permitted. In this case, at the population level, an individual with covariate X_* is classified as $Y = 1$ if $c_L > 0$, as $Y = 0$ if $c_U < 0$, and is classified at random (with equal probability) if $c_L \leq 0 \leq c_U$. In the sample, classification follows the same rule using the same randomization device. As before, a classification error occurs whenever the sample classification deviates from the population classification. In other words, while the definition of population classification differs from that in Panel A, the definition of classification error remains unchanged. The theoretical analysis relevant to this case is presented in Section 5.5. These simulation results are consistent with the findings in Section 5.5, in the sense that misclassification errors are smaller in magnitude, as implied by Theorem 5.3.

Panel C in Table 4 compares the population classification rule—identical to that in Panel B, using the same randomization device—with a sample classification rule based solely on the sign of the sample frequency of $f(x) = \mathbb{P}(Y = 1 \mid X = x) - \tau$. The resulting classification errors are large and exhibit little to no improvement as n increases. This is because the sample frequencies of $f(x)$ are already estimated with high precision at $n = 2880$, leaving limited scope for further gains. These results underscore the value of extrapolation through the maximum score model, even when data are available in every cell. The advantages will become even more pronounced in settings with sparse or missing observations.

10. Conclusions

We have shown that scalable inference based on linear programming can be conducted for maximum score estimation. We have demonstrated the usefulness of our proposal of credible binary classification rules. In our empirical examples, the finite sample inference methods did not provide informative bounds but the asymptotic inference methods did. One natural follow-up question is whether it would be possible to improve the confidence regions constructed via the Bonferroni correction. This is an interesting topic for future research.

Appendix A. Proofs

Proof of Theorem 4.1. We divide the analysis into three cases based on the identified interval $[c_L(x), c_U(x)]$:

(1) **Case 1:** $c_L(x) > 0$**.**

If $c_L(x) > 0$, then every admissible $\beta' x$ in $[c_L(x), c_U(x)]$ is strictly positive.

Thus, predicting 1 incurs the smallest cost C_b, while predicting 0 could incur the larger cost C, and abstaining costs $C_\emptyset$, which is strictly greater than C_b. Hence, $\hat{y} = 1$ is optimal.

(2) **Case 2:** $c_U(x) < 0$**.**
If $c_U(x) < 0$, then $\beta' x < 0$ for all feasible values. Predicting 0 again incurs C_b, while predicting 1 could cost C, and abstaining costs $C_\emptyset > C_b$. So $\hat{y} = 0$ is optimal.

(3) **Case 3:** $0 \in [c_L(x), c_U(x)]$**.**
Here, $\beta' x$ could be either positive or negative. If we predict $\hat{y} = 1$ but $\beta' x < 0$, we incur cost C; similarly, predicting $\hat{y} = 0$ when $\beta' x > 0$ also results in cost C. Abstaining ($\hat{y} = \emptyset$) leads to a smaller cost $C_\emptyset < C$, regardless of the true sign. Therefore, abstaining is optimal in a *worst-case* sense.

Combining the three cases yields the minimax classification rule stated above. □

Proof of Theorem 4.2. As in the proof of Theorem 4.1, we divide the analysis into three cases based on the identified interval $[c_L(x), c_U(x)]$. The first two cases remain the same.

Now consider Case 3, in which $[c_L(x), c_U(x)]$ contains 0. In this case, suppose the classification is randomized such that the observation is assigned $Y = 1$ with probability p and $Y = 0$ with probability $1 - p$. The corresponding expected loss is given in (4.5) and the associated regret is described in (4.6). Hence, the maximum regret is $(C - C_b)\max(p, 1 - p)$. Minimizing this expression yields $p = 0.5$. As $(C - C_b)/2$ is strictly greater than C_b, it is optimal to randomize with equal probability. □

Proof of Theorem 5.1. Suppose that $g_0 \in \mathcal{G}_\alpha$. Any feasible solution $(r'b)$ of (3.4)-(3.5) is also a feasible solution $(r'b)$ of (5.2)-(5.3). Therefore, the feasible region of (5.2)-(5.3) contains the feasible region of (3.4)-(3.5). Consequently, $\hat{c}_L \leq c_L \leq c_U \leq \hat{c}_U$, which proves the theorem. □

Proof of Theorem 5.2. We focus on the case where X_* is random; the fixed X_* case follows similarly.

Under (5.1), combined with the independence of X_* from the estimation sample, we have almost surely:

$$\mathbb{P}\left(\mathcal{C}(X_*) = 1 \mid X_*\right) \geq 1 - \alpha, \tag{A.1}$$

where

$$\mathcal{C}(X_*) = \begin{cases} 1 & \text{if } \hat{c}_L(X_*) \leq c_L(X_*) \leq \beta' X_* \leq c_U(X_*) \leq \hat{c}_U(X_*), \\ 0 & \text{otherwise.} \end{cases}$$

Decompose the misclassification probability:

$$\begin{aligned} \mathbb{P}(\mathcal{M}(X_*) = 1) &= \mathbb{P}(\mathcal{M}(X_*) = 1, \mathcal{C}(X_*) = 0) + \mathbb{P}(\mathcal{M}(X_*) = 1, \mathcal{C}(X_*) = 1) \\ &\leq \mathbb{P}(\mathcal{C}(X_*) = 0) + \mathbb{P}(\mathcal{M}(X_*) = 1, \mathcal{C}(X_*) = 1) \\ &\leq \alpha + \mathbb{P}(\mathcal{M}(X_*) = 1, \mathcal{C}(X_*) = 1), \end{aligned}$$

where the last inequality follows from (A.1).

It remains to bound the second term above. To simply the notation, we will suppress dependence on X_* in the remainder of the proof. Misclassification can happen in three distinct cases.

- Case 1: Oracle classifies $Y = 1$ $(c_L > 0)$

  The oracle classifies as $Y = 1$. Misclassification occurs if $\hat{c}_L \leq 0$. Thus,

  $$\mathbb{P}(\mathcal{M} = 1, \mathcal{C} = 1, c_L > 0) \leq \mathbb{P}\left(\hat{c}_L \leq 0, c_L > 0\right).$$

- Case 2: Oracle classifies $Y = 0$ $(c_U < 0)$

  The oracle classifies as $Y = 0$. Misclassification occurs if $\hat{c}_U \geq 0$. Thus,

  $$\mathbb{P}(\mathcal{M} = 1, \mathcal{C} = 1, c_U < 0) \leq \mathbb{P}\left(\hat{c}_U \geq 0, c_U < 0\right).$$

- Case 3: Oracle does not classify $(c_L \leq 0 \leq c_U)$

  The oracle abstains from classification. Misclassification occurs if either $\hat{c}_L > 0$ or $\hat{c}_U < 0$. Using the union bound,

  $$\begin{aligned} &\mathbb{P}(\mathcal{M} = 1, \mathcal{C} = 1, c_L \leq 0 \leq c_U) \\ &\leq \mathbb{P}\left(\hat{c}_L > 0, \hat{c}_L \leq c_L \leq 0\right) + \mathbb{P}\left(\hat{c}_U < 0, \hat{c}_U \geq c_U \geq 0\right) \\ &= 0, \end{aligned}$$

  where the last equality follows from the simple fact that the relevant events cannot happen.

Combining the three cases yields

$$\mathbb{P}(\mathcal{M} = 1, \mathcal{C} = 1) \leq \mathbb{P}\left(\hat{c}_L \leq 0, c_L > 0\right) + \mathbb{P}\left(\hat{c}_U \geq 0, c_U < 0\right).$$

We further write:

$$\begin{aligned}\mathbb{P}\left(\hat{c}_L \leq 0, c_L > 0\right) &= \mathbb{P}(\hat{c}_L \leq 0, 0 < c_L \leq \epsilon) + \mathbb{P}(\hat{c}_L \leq 0, c_L > \epsilon) \\ &\leq \mathbb{P}(0 < c_L \leq \epsilon) + \mathbb{P}(\hat{c}_L - c_L \leq -\epsilon) \\ &\leq M_L\epsilon + \mathbb{P}(\hat{c}_L - c_L \leq -\epsilon).\end{aligned}$$

where the last inequality comes from Assumption 4 (b) (ii). Similarly,

$$\begin{aligned}\mathbb{P}(\hat{c}_U \geq 0, c_U < 0) &= \mathbb{P}(\hat{c}_U \geq 0, -\epsilon \leq c_U < 0) + \mathbb{P}(\hat{c}_U \geq 0, -\epsilon > c_U) \\ &\leq \mathbb{P}(-\epsilon \leq c_U < 0) + \mathbb{P}(\hat{c}_U - c_U \geq \epsilon) \\ &\leq M_U\epsilon + \mathbb{P}(\hat{c}_U - c_U \geq \epsilon),\end{aligned}$$

where again the last inequality comes from Assumption 4 (b) (ii). Substituting these into the earlier decomposition gives the desired conclusion. □

Proof of Theorem 5.3. As in the proof of Theorem 5.2, we focus on the case where X_* is random; the fixed design case follows similarly. As before,

$$\mathbb{P}(\mathcal{M}(X_*) = 1) \leq \alpha + \mathbb{P}(\mathcal{M}(X_*) = 1, \mathcal{C}(X_*) = 1).$$

To simplify notation, we suppress the dependence on X_* in what follows. Misclassification can occur in the following cases:

Case 1: Oracle classifies $Y = 1$ ($c_L > 0$). The maximum score rule classifies correctly if $\hat{c}_L > 0$; otherwise, it randomly classifies $Y = 1$ or $Y = 0$. Hence,

$$\begin{aligned}\mathbb{P}(\mathcal{M} = 1, \mathcal{C} = 1, c_L > 0) &= \mathbb{P}\left(r = 0, \hat{c}_L \leq 0, c_L > 0, \mathcal{C} = 1\right) \\ &\leq 0.5\mathbb{P}\left(\hat{c}_L \leq 0, c_L > 0\right).\end{aligned}$$

Case 2: Oracle classifies $Y = 0$ ($c_U < 0$). The maximum score rule classifies correctly if $\hat{c}_U < 0$; otherwise, it randomly classifies $Y = 1$ or $Y = 0$. Hence,

$$\begin{aligned}\mathbb{P}(\mathcal{M} = 1, \mathcal{C} = 1, c_U < 0) &= \mathbb{P}\left(r = 0, \hat{c}_U \geq 0, c_U < 0, \mathcal{C} = 1\right) \\ &\leq 0.5\mathbb{P}\left(\hat{c}_U \geq 0, c_U < 0\right).\end{aligned}$$

Case 3: Oracle randomizes ($c_L \leq 0 \leq c_U$). Under $\mathcal{C} = 1$, we have $\hat{c}_L \leq c_L < c_U \leq \hat{c}_U$, so both oracle and maximum score classifiers use the same random classification r. Therefore,

$$M_{\mathrm{OR}} = M_{\mathrm{MS}}, \quad \text{and} \quad \mathbb{P}(\mathcal{M} = 1, \mathcal{C} = 1, c_L \leq 0 \leq c_U) = 0.$$

Combining the cases, we obtain:

$$\mathbb{P}(\mathcal{M}=1,\mathcal{C}=1) \leq 0.5\,\mathbb{P}(\hat{c}_L \leq 0, c_L > 0) + 0.5\,\mathbb{P}(\hat{c}_U \geq 0, c_U < 0).$$

The rest of the proof proceeds identically to that of Theorem 5.2, yielding the stated bounds. □

Proof of Theorem 6.1. Note that it only matters whether g_j is positive or negative in the bilinear constraints (5.3). Therefore, when $\mathcal{G}_\alpha$ is a Cartesian product of intervals as in (6.1), solving (5.2)-(5.3) is equivalent to solving

$$\underset{b\in\mathbf{B};\, g\in\mathcal{G}_\alpha}{\text{maximize}} \left(\underset{b\in\mathbf{B};\, g\in\mathcal{G}_\alpha}{\text{minimize}}\right) :\ r'b \tag{A.2}$$

subject to (6.3) and

$$g_j\left(x_{j,1} + b_2 x_{j,2} + \cdots + b_q x_{j,q}\right) \geq 0 \ \text{ if } -\hat{s}(x_j,\alpha) \leq \hat{g}(x_j) \leq \hat{s}(x_j,\alpha). \tag{A.3}$$

Suppose that we want to solve the maximization problem in (A.2) under (6.3) and (A.3). We first start with the maximum under (6.3) only. We now verify that the same maximum is achievable under (6.3) and (A.3). Let $\tilde{b}_{LP}$ denote a maximizer under the relaxed LP problem (that is, a solution under (6.3) only). If we add additional restrictions in (A.3), the resulting maximum cannot be larger and so $r'\tilde{b}_{LP}$ will be the maximum value under (6.3) and (A.3), if we can verify that $\tilde{b}_{LP}$ is feasible under (A.3). If $x_j'\tilde{b}_{LP} \geq 0$, we can choose a $g_j > 0$ as a solution; If $x_j'\tilde{b}_{LP} < 0$, we can choose a different $g_j < 0$. This is possible because both positive and negative values of g_j are feasible under (A.3). The case of the minimization problem can be dealt in a similar way. Thus, we have obtained the desired result. □

Proof of Theorem 7.1. Let $W \equiv (X, V_0, V_1)$ denote the observed covariates and $w \equiv (x, v_0, v_1)$. By the law of iterated expectations and assumptions M and MI,

$$\begin{aligned}
\mathbb{E}\left(Y \mid W = w\right) &= \int_{v_0}^{v_1} \mathbb{E}\left(Y \mid W = w, V\right) d\mathbb{P}\left(V \mid W = w\right) \\
&= \int_{v_0}^{v_1} \mathbb{E}\left(Y \mid X = x, V\right) d\mathbb{P}\left(V \mid W = w\right) \\
&\leq \mathbb{E}\left(Y \mid X = x, V = v_1\right),
\end{aligned}$$

which establishes the second inequality in (7.1). Analogously,

$$\begin{aligned}\mathbb{E}\left(Y \mid W=w\right) &= \int_{v_0}^{v_1} \mathbb{E}\left(Y \mid W=w, V\right) d\mathbb{P}\left(V \mid W=w\right) \\ &= \int_{v_0}^{v_1} \mathbb{E}\left(Y \mid X=x, V\right) d\mathbb{P}\left(V \mid W=w\right) \\ &\geq \mathbb{E}\left(Y \mid X=x, V=v_0\right),\end{aligned}$$

which establishes the first inequality in (7.1).

To see why the bounds are sharp, suppose that $\mathbb{E}\left(Y \mid W=w\right) \leq \mathbb{E}\left(Y \mid X=x, V=v_1\right)$ is not sharp. Then there is a constant κ such that

$$\mathbb{E}\left(Y \mid W=w\right) = \int_{v_0}^{v_1} \mathbb{E}\left(Y \mid X=x, V\right) d\mathbb{P}\left(V \mid W=w\right) < \kappa \leq \mathbb{E}\left(Y \mid X=x, V=v_1\right).$$

By M, the existence of a κ satisfying the strict inequality requires extending the limits of integration to $v_0 - \zeta_-$ and/or $v_1 + \zeta_+$, where $\zeta_-, \zeta_+ \geq 0$ are constants and at least one is strictly positive. Under assumption I, this does not change the value of the integral, so the strict inequality above is impossible. If it were possible, the resulting κ would exceed $\mathbb{E}\left(Y \mid X=x, V=v_1\right)$. A similar argument applies to the other bound. □

Appendix B. Comparison of Finite and Asymptotic Inference: Monte Carlo Results

This section presents the results of small Monte Carlo experiments to compare the finite and asymptotic inference methods. Specifically, the binary $Y_i \in \{0,1\}$ is generated from

$$\begin{aligned}Y_i &= I(Y_i^* \geq 0), \\ Y_i^* &= X_i'\beta_0 + U_i,\end{aligned}$$

where $X_i = (1, X_{1i}, X_{2i})'$, $U_i = \sigma(X_i)V_i$, and $V_i \sim N(0,1)$. Here, the covariates are generated as follows. First, W_{1i} and W_{2i} are generated from bivariate normals with means of zero, variances of one, and covariance 0.25. For each $k = 1, 2$, $X_{ki} = -2$ if $W_{ki} \leq \Phi^{-1}(0.2)$; $X_{ki} = -1$ if $\Phi^{-1}(0.2) < W_{ki} \leq \Phi^{-1}(0.4)$; $X_{ki} = 0$ if $\Phi^{-1}(0.4) < W_{ki} \leq \Phi^{-1}(0.6)$; $X_{ki} = 1$ if $\Phi^{-1}(0.6) < W_{ki} \leq \Phi^{-1}(0.8)$; $X_{ki} = 2$ if $\Phi^{-1}(0.8) < W_{ki}$. In words, covariates are discretized by equal quantile binning. Here, $J = 25$. We set $\beta_0 = (\beta_0^{(0)}, \beta_0^{(1)}, \beta_0^{(2)}) = (0.5, 1, 2)'$ and consider the heteroskedastic design where

$\sigma(X_i) = 0.15[1 + (X_{1i} + X_{2i})^2]$. In this design, population values of $g_0(x_j)$ are never zero. The confidence level is set at $\alpha = 0.05$.

B.1. **Non-Asymptotic Inference.** We first consider non-asymptotic inference. The sample size was $n = \{5000, 10000, 15000, 20000, 25000\}$, and the number of replications was 100 for each experiment. The range of sample sizes enables the experiments to illustrate the looseness of the non-asymptotic bounds with smaller sample sizes and the effect of sample size on the estimated bounds. In solving the linear programming problem, each component of $b$was assumed to be between $[-10, 10]$.

Table 5 summarizes the results of Monte Carlo experiments, focusing on $\beta_0^{(2)}$, whose bounds are $[1.5, 3]$. The mean and standard deviation of each estimated bound are shown in the table. The coverage refers to the proportion of estimated bounds that contain the population bounds. The latter is 1 for all cases. The estimation results show that (i) the coverage probability is 1 for all cases; (ii) when $n = 25000$, the estimated bounds match the population bounds in all simulation draws, meaning the population bounds are estimated without any error; and (iii) as expected, the bounds are loose unless n is large enough.

TABLE 5. Results of Monte Carlo Experiments: Non-Asymptotic Inference

$\beta_0^{(2)}$ — Population Bounds $[1.5, 3]$					
Sample	Mean		Std. Dev.		Coverage
Size	Lower Bound	Upper Bound	Lower Bound	Upper Bound	
5000	-1.038	10.000	0.504	0.000	1.000
10000	0.740	10.000	0.345	0.000	1.000
15000	1.065	9.370	0.169	2.013	1.000
20000	1.475	3.280	0.110	1.379	1.000
25000	1.500	3.000	0.000	0.000	1.000

B.2. **Asymptotic Inference.** We now consider asymptotic inference. The sample sizes were $n = \{500, 750, 1000, 2000\}$, and the number of simulations was 1000 for each experiment. As before, in solving the linear programming problems, each component of b was assumed to be between $[-10, 10]$. Table 6 summarizes the results of Monte Carlo experiments, focusing on $\beta_0^{(2)}$, whose population bounds are $[1.5, 3]$. Not surprisingly, the estimated bounds are much tighter with asymptotic inference. When $n = 2000$, the true population bound is obtained in each simulation draw.

TABLE 6. Results of Monte Carlo Experiments: Asymptotic Inference

$\beta_0^{(2)}$ — Population Bounds [1.5, 3]					
Sample	Mean		Std. Dev.		Coverage
Size	Lower Bound	Upper Bound	Lower Bound	Upper Bound	
500	1.394	4.617	0.204	2.894	1.000
750	1.488	3.205	0.075	1.176	1.000
1000	1.500	3.000	0.016	0.000	1.000
2000	1.500	3.000	0.000	0.000	1.000

APPENDIX C. CLUSTERED DEPENDENCE AND SURVEY SAMPLING WEIGHTS

In applications, it is often the case that observations are not necessarily independent but can be dependent within a cluster. Furthermore, it is common to have non-uniform survey sampling weights. In fact, our second empirical example in Section 8.3 has these features. In this section, we extend our framework to accommodate clustered dependence as well as survey sampling weights.

We begin by redefining $\hat{g}(x_j)$ and $g_0(x_j)$. First, for the fixed design, let $Y_{\ell,j,c}$ and $w_{\ell,j,c}$, respectively, denote the binary outcome and survey sampling weight for individual ℓ with covariate x_j in cluster c. Define

$$N_j \equiv \sum_{c=1}^{m_j} \sum_{\ell=1}^{n_{j,c}} w_{\ell,j,c}, \quad \Upsilon_{j,c} \equiv \sum_{\ell=1}^{n_{j,c}} w_{\ell,j,c}(Y_{\ell,j,c} - \tau), \quad \text{and} \quad \overline{\Upsilon}_j \equiv N_j^{-1} \sum_{c=1}^{m_j} \Upsilon_{j,c},$$

where $n_{j,c}$ is the number of observations with $X = x_j$ in cluster c and m_j is the number of clusters with $X = x_j$. Abusing notation a bit, redefine

$$\begin{aligned} p(x_j) &\equiv \frac{N_j}{\sum_{j=1}^{J} N_j}, \\ \hat{g}\,(x_j) &\equiv \frac{\sum_{c=1}^{m_j} \Upsilon_{j,c}}{\sum_{j=1}^{J} N_j} = \overline{\Upsilon}_j\, p(x_j), \\ g_0(x_j) &\equiv \mathbb{E}(\overline{\Upsilon}_j) p(x_j). \end{aligned}$$

In the random design case, let $(Y_{i,c}, X_{i,c}, w_{i,c})$, respectively, denote the binary outcome, covariates, and sampling weight for individual i in cluster c. Also, let n_c denote the number of observations in cluster c, m the number of clusters and

$N \equiv \sum_{c=1}^{m} \sum_{i=1}^{n_c} w_{i,c}$. Redefine

$$\begin{aligned}\hat{g}\left(x_j\right) &\equiv \frac{\sum_{c=1}^{m} \sum_{i=1}^{n_c} w_{i,c}(Y_{i,c} - \tau) I\left(X_{i,c} = x_j\right)}{\sum_{c=1}^{m} \sum_{i=1}^{n_c} w_{i,c}} \\ &= N^{-1} \sum_{c=1}^{m} \Upsilon_{j,c}, \quad \text{and} \\ g_0(x_j) &\equiv N^{-1} \sum_{c=1}^{m} \mathbb{E}(\Upsilon_{j,c}),\end{aligned}$$

where

$$\Upsilon_{j,c} \equiv \sum_{i=1}^{n_c} w_{i,c}(Y_{i,c} - \tau) I\left(X_{i,c} = x_j\right).$$

Note that in both fixed and random designs, $\hat{g}\left(x_j\right)$ is an unbiased estimator of $g_0(x_j)$ for each x_j.

To develop inference methods, we make the following additional assumption.

Assumption 8. *(i) Observations are independent across different clusters. (ii) Survey sampling weights are not random but fixed for both fixed and random designs.*

Part (i) of Assumption 8 is standard and part (ii) assumes that the sampling weights are fixed in both designs.

C.1. **Finite Sample Inference.**

C.1.1. *Fixed Design Case.* We modify the confidence region in Section 5.2.1 as follows. Observe that

$$-\tau \gamma_{j,c} \leq N_j^{-1} \Upsilon_{j,c} \leq (1 - \tau) \gamma_{j,c},$$

where

$$\gamma_{j,c} \equiv \frac{1}{N_j} \sum_{\ell=1}^{n_{j,c}} w_{\ell,j,c}.$$

Under Assumption 8, an application of Hoeffding's inequality for the sum of $N_j^{-1} \Upsilon_{j,c}$ over c yields that

$$\mathbb{P}\left[\left|\hat{g}\left(x_j\right) - g_0\left(x_j\right)\right| \geq p\left(x_j\right) t_j\right] \leq 2 \exp\left(-2 \frac{t_j^2}{\sum_{c=1}^{m_j} \gamma_{j,c}^2}\right)$$

for any $t_j > 0$ and each $j = 1, \ldots, J$. Moreover,

$$\hat{g}(x_j) - p(x_j)\, t_j \leq g_0(x_j) \leq \hat{g}(x_j) + p(x_j)\, t_j$$

for all $j = 1, \ldots, J$ with probability at least $1 - 2\sum_{j=1}^{J} \exp\left(-2t_j^2 / \sum_{c=1}^{m_j} \gamma_{j,c}^2\right)$. Set

$$t_j = \left(\frac{\sum_{c=1}^{m_j} \gamma_{j,c}^2}{2} \log \frac{2J}{\alpha} \right)^{1/2} \equiv t_{j,\alpha}^*.$$

Modify the confidence set in (5.6) with the new critical value $t_{j,\alpha}^*$. That is, we now set

$$\text{(C.1)} \qquad \mathcal{G}_\alpha = \left\{ (g_1, \ldots, g_J) : \hat{g}(x_j) - p(x_j)\, t_{j,\alpha}^* \leq g_j \leq \hat{g}(x_j) + p(x_j)\, t_{j,\alpha}^* \ \forall j \right\}.$$

Then $g \in \mathcal{G}_\alpha$ holds with probability at least $1 - \alpha$. In view of (5.5), the quantity $[\sum_{c=1}^{m_j} \gamma_{j,c}^2]^{-1}$ can be viewed as the *effective* sample size. If the survey sampling weights are uniform and $n_{j,c}$ are identical across clusters, then it reduces to m_j, that is, the number of clusters for the $(X = x_j)$ observations.

C.1.2. *Random Design Case.* We extend the confidence region in Section 5.2.2 as follows. Note that

$$-\tau\gamma_c \leq N^{-1}\Upsilon_{j,c} \leq (1-\tau)\gamma_c,$$

where

$$\gamma_c \equiv \frac{1}{N} \sum_{i=1}^{n_c} w_{i,c}.$$

As in the previous subsection, under Assumption 8, another application of Hoeffding's inequality yields that

$$\mathbb{P}\left[|\hat{g}(x_j) - g_0(x_j)| \geq t \right] \leq 2\exp\left(-2\frac{t^2}{\sum_{c=1}^{m} \gamma_c^2} \right)$$

for any $t > 0$ and each $j = 1, \ldots, J$. Moreover,

$$\hat{g}(x_j) - t \leq g_0(x_j) \leq \hat{g}(x_j) + t$$

for all $j = 1, \ldots, J$ with probability at least $1 - 2J\exp\left(-2t^2 / \sum_{c=1}^{m} \gamma_c^2\right)$. Set

$$t = \left(\frac{\sum_{c=1}^{m} \gamma_c^2}{2} \log \frac{2J}{\alpha} \right)^{1/2} \equiv t_\alpha^*.$$

Modify the confidence set in (5.8) with the new critical value t_α^*. That is, we now set

$$\mathcal{G}_\alpha = \{(g_1, \ldots, g_J) : \hat{g}(x_j) - t_\alpha^* \leq g_j \leq \hat{g}(x_j) + t_\alpha^* \ \forall j\}. \tag{C.2}$$

Then $g \in \mathcal{G}_\alpha$ holds with probability at least $1 - \alpha$. As in the previous section, the quantity $[\sum_{c=1}^m \gamma_c^2]^{-1}$ can be viewed as the *effective* sample size. As in Section 5.2.1, the confidence region under the random design case will be larger if

$$p(x_j)\, t_j < t_\alpha^*$$

equivalently,

$$N_j \left(\sum_{c=1}^{m_j} \gamma_{j,c}^2 \right)^{1/2} < \left(\sum_{j=1}^{J} N_j \right) \left(\sum_{c=1}^{m} \gamma_c^2 \right)^{1/2}.$$

C.2. **Asymptotic Inference.**

C.2.1. *Fixed Design Case.* Note that the variance of $(N_j)^{1/2}(\overline{\Upsilon}_j - \mathbb{E}\overline{\Upsilon}_j)$ is

$$\mathbb{V}_j \equiv \mathbb{E}\left[N_j (\overline{\Upsilon}_j - \mathbb{E}\overline{\Upsilon}_j)^2 \right] = N_j^{-1} \sum_{c=1}^{m_j} \mathbb{V}_{j,c},$$

where

$$\begin{aligned} \mathbb{V}_{j,c} &\equiv N_j^{-1} \mathbb{E}\left[\{\Upsilon_{j,c} - \mathbb{E}(\Upsilon_{j,c})\}^2 \right] \\ &= \frac{1}{N_j} \sum_{\ell=1}^{n_{j,c}} \sum_{k=1}^{n_{j,c}} w_{\ell,j,c} w_{k,j,c} \mathbb{E}\left[\{Y_{\ell,j,c} - \mathbb{E}(Y_{\ell,j,c})\} \{Y_{k,j,c} - \mathbb{E}(Y_{k,j,c})\} \right]. \end{aligned}$$

Define $\mathbb{V}_j \equiv \sum_{c=1}^{m_j} \mathbb{V}_{j,c}$. To apply the Lyapunov central limit theorem, we make the following assumption.

Assumption 9. *For each j and some $\delta_j > 0$,*

$$m_j^{-1} \mathbb{V}_j^{1+\delta/2} \to \infty.$$

As $Y_{\ell,j,c}$ is a bounded random variable, the Lyapunov condition is satisfied under Assumption 9. To appreciate Assumption 9, consider a special case such that $w_{\ell,j,c} \equiv 1$ and the covariance between $Y_{\ell,j,c}$ and $Y_{k,j,c}$ is uniformly bounded from below by a constant $c_j > 0$. Then,

$$\mathbb{V}_j \geq c_j \frac{\sum_{c=1}^{m_j} n_{j,c}^2}{\sum_{c=1}^{m_j} n_{j,c}} \geq c_j \min_{c=1,\ldots,m_j} n_{j,c},$$

which indicates that Assumption 9 is satisfied if $\min_{c=1,\ldots,m_j} n_{j,c}^{1+\delta/2}/m_j \to \infty$ for some $\delta > 0$. It is not strictly necessary to assume Assumption 9; alternatively, one may

invoke a more general condition for the Lindeberg-Feller central limit theorem. By the Lyapunov central limit theorem, we have that

$$\mathbb{V}_j^{-1/2}(N_j)^{1/2}\left[\overline{\Upsilon}_j - \mathbb{E}(\overline{\Upsilon}_j)\right] \to_d N(0,1).$$

To develop an asymptotically valid inference method, write

$$\mathbb{V}_{j,c} = \mathbb{V}_{1,j,c} - \mathbb{V}_{2,j,c},$$

where

$$\mathbb{V}_{1,j,c} \equiv N_j^{-1}\sum_{\ell=1}^{n_{j,c}}\sum_{k=1}^{n_{j,c}} w_{\ell,j,c}w_{k,j,c}\mathbb{E}\left[Y_{\ell,j,c}Y_{k,j,c}\right],$$
$$\mathbb{V}_{2,j,c} \equiv N_j^{-1}\sum_{\ell=1}^{n_{j,c}}\sum_{k=1}^{n_{j,c}} w_{\ell,j,c}w_{k,j,c}\mathbb{E}(Y_{\ell,j,c})\mathbb{E}(Y_{k,j,c}).$$

To estimate the first term $\mathbb{V}_{1,j,c}$, observe that an unbiased estimator of $\mathbb{E}\left[Y_{\ell,j,c}Y_{k,j,c}\right]$ is

$$Y_{\ell,j,c}Y_{k,j,c},$$

which suggests that we estimate $\mathbb{V}_{1,j,c}$ by

$$\hat{\mathbb{V}}_{1,j,c} \equiv N_j^{-1}\sum_{\ell=1}^{n_{j,c}}\sum_{k=1}^{n_{j,c}} w_{\ell,j,c}w_{k,j,c}Y_{\ell,j,c}Y_{k,j,c}.$$

To estimate the second term, it is necessary to estimate $\mathbb{E}(Y_{\ell,j,c})\mathbb{E}(Y_{k,j,c})$ consistently, which is more challenging than coming up with an unbiased estimator of $\mathbb{E}\left[Y_{\ell,j,c}Y_{k,j,c}\right]$. Furthermore, if it is not well estimated, the resulting estimator of $\mathbb{V}_{j,c}$ can be negative in finite samples. In view of this, we aim to estimate an upper bound on $\mathbb{V}_{j,c}$ by dropping $\mathbb{V}_{2,j,c}$.

Let $\hat{\mathbb{V}}_{1,j} \equiv \sum_{c=1}^{m_j}\hat{\mathbb{V}}_{1,j,c}$ and we assume the following condition.

Assumption 10. *For each j, $\mathbb{P}\left(\hat{\mathbb{V}}_{1,j} \geq \mathbb{V}_j\right) \to 1$.*

This assumption ensures that the following confidence region is asymptotically valid:

(C.3)

$$\mathcal{G}_\alpha = \left\{(g_1,\ldots,g_J) : \hat{g}(x_j) - \frac{N_j^{1/2}\hat{\mathbb{V}}_{1,j}^{1/2}}{N}z_{1-\alpha/(2J)} \leq g_j \leq \hat{g}(x_j) + \frac{N_j^{1/2}\hat{\mathbb{V}}_{1,j}^{1/2}}{N}z_{1-\alpha/(2J)}\ \forall j\right\},$$

where $N \equiv \sum_{j=1}^{J} N_j$.

C.2.2. *Random Design Case.* Define

$$\Xi_{i,c}(x_j) \equiv (Y_{i,c} - \tau) I\left(X_{i,c} = x_j\right) - \mathbb{E}\left[(Y_{i,c} - \tau) I\left(X_{i,c} = x_j\right)\right].$$

Note that the variance of $N^{1/2}[\hat{g}\left(x_j\right) - g_0\left(x_j\right)]$ is

$$\mathbb{V}_j \equiv \sum_{c=1}^{m} \mathbb{V}_{j,c},$$

where

$$\begin{aligned}\mathbb{V}_{j,c} &\equiv \frac{1}{N}\mathbb{E}[\{\Upsilon_{j,c} - \mathbb{E}(\Upsilon_{j,c})\}^2] \\ &= \frac{1}{N}\sum_{i=1}^{n_c}\sum_{k=1}^{n_c} w_{i,c} w_{k,c} \mathbb{E}\left[\Xi_{i,c}(x_j)\Xi_{k,c}(x_j)\right].\end{aligned}$$

As in the previous subsection, assume that $m^{-1}\mathbb{V}_j^{1+\delta} \to \infty$ for some $\delta > 0$ and invoke the Lyapunov central limit theorem to obtain

$$\mathbb{V}_j^{-1/2} N^{1/2}\left[\hat{g}\left(x_j\right) - g_0\left(x_j\right)\right] \to_d N(0,1).$$

As before, estimate the upper bound on $\mathbb{V}_{j,c}$ by

$$\hat{\mathbb{V}}_{1,j,c} \equiv \frac{1}{N}\sum_{i=1}^{n_c}\sum_{k=1}^{n_c} w_{i,c} w_{k,c} (Y_{i,c} - \tau) I\left(X_{i,c} = x_j\right)(Y_{k,c} - \tau) I\left(X_{k,c} = x_j\right)$$

and $\hat{\mathbb{V}}_{1,j} \equiv \sum_{c=1}^{m} \hat{\mathbb{V}}_{1,j,c}$. Assume again that for each j, $\mathbb{P}\left(\hat{\mathbb{V}}_{1,j} \geq \mathbb{V}_j\right) \to 1$. Finally, we modify (5.10) as follows:

(C.4)

$$\mathcal{G}_\alpha = \left\{(g_1, \ldots, g_J) : \hat{g}(x_j) - \frac{\hat{\mathbb{V}}_{1,j}^{1/2}}{N^{1/2}} z_{1-\alpha/(2J)} \leq g_j \leq \hat{g}(x_j) + \frac{\hat{\mathbb{V}}_{1,j}^{1/2}}{N^{1/2}} z_{1-\alpha/(2J)} \ \forall j\right\}.$$

It is interesting to note that unlike the previous settings, the confidence regions are identical between the fixed and random design cases here. That is, $\mathcal{G}_\alpha$ in (C.3) is identical to $\mathcal{G}_\alpha$ in (C.4).

References

Abrevaya, J., and J. Huang (2005): "On the bootstrap of the maximum score estimator," *Econometrica*, 73(4), 1175–1204.

Benoit, D. F., and D. Van den Poel (2012): "Binary quantile regression: a Bayesian approach based on the asymmetric Laplace distribution," *Journal of Applied Econometrics*, 27(7), 1174–1188.

Blevins, J. R. (2015): "Non-Standard Rates of Convergence of Criterion-Function-Based Set Estimators," *Econometrics Journal*, 18, 172–199.

Breza, E., A. G. Chandrasekhar, and D. Viviano (2025): "Generalizability with ignorance in mind: learning what we do (not) know for archetypes discovery," arXiv:2501.13355 [econ.EM] `https://arxiv.org/abs/2501.13355`.

Cattaneo, M. D., M. Jansson, and K. Nagasawa (2020): "Bootstrap-Based Inference for Cube Root Asymptotics," *Econometrica*, 88(5), 2203–2219.

Chen, L.-Y., and S. Lee (2018): "Best subset binary prediction," *Journal of Econometrics*, 206(1), 39–56.

——— (2019): "Breaking the curse of dimensionality in conditional moment inequalities for discrete choice models," *Journal of Econometrics*, 210(2), 482–497.

——— (2020): "Binary classification with covariate selection through ℓ_0-penalised empirical risk minimisation," *Econometrics Journal*, 24(1), 103–120.

Corno, L., N. Hildebrandt, and A. Voena (2020): "Age of Marriage, Weather Shocks, and the Direction of Marriage Payments," *Econometrica*, 88(3), 879–915.

Delgado, M. A., J. M. Rodriguez-Poo, and M. Wolf (2001): "Subsampling inference in cube root asymptotics with an application to Manski's maximum score estimator," *Economics Letters*, 73(2), 241–250.

Dempster, A. (2008): "The Dempster–Shafer calculus for statisticians," *International Journal of Approximate Reasoning*, 48(2), 365–377.

El-Yaniv, R., and Y. Wiener (2010): "On the Foundations of Noise-free Selective Classification," *Journal of Machine Learning Research*, 11(53), 1605–1641.

Florios, K., and S. Skouras (2008): "Exact computation of max weighted score estimators," *Journal of Econometrics*, 146(1), 86–91.

Gao, W. Y., S. Xu, and K. Xu (2022): "Two-Stage Maximum Score Estimator," arXiv:2009.02854 [econ.EM], `https://arxiv.org/abs/2009.02854`.

Hendrickx, K., L. Perini, D. Van der Plas, W. Meert, and J. Davis (2024): "Machine learning with a reject option: A survey," *Machine Learning*, 113(5), 3073–3110.

Horowitz, J. L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60(3), 505–531.

——— (2002): "Bootstrap critical values for tests based on the smoothed maximum score estimator," *Journal of Econometrics*, 111(2), 141–167.

Horowitz, J. L., and S. Lee (2023): "Inference in a Class of Optimization Problems: Confidence Regions and Finite Sample Bounds on Errors in Coverage Probabilities," *Journal of Business & Economic Statistics*, 41(3), 927–938.

Johnson, D., and F. Preparata (1978): "The densest hemisphere problem," *Theoretical Computer Science*, 6(1), 93–107.

Jun, S. J., J. Pinkse, and Y. Wan (2015): "Classical Laplace estimation for $\sqrt[3]{n}$-consistent estimators: improved convergence rates and rate-adaptive inference," *Journal of Econometrics*, 187(1), 201–216.

——— (2017): "Integrated Score Estimation," *Econometric Theory*, 33(6), 1418–56.

Kessler, J. B., C. Low, and C. D. Sullivan (2019): "Incentivized Resume Rating: Eliciting Employer Preferences without Deception," *American Economic Review*, 109(11), 3713–44.

Khan, S., T. Komarova, and D. Nekipelov (2024): "Sharp and Robust Estimation of Partially Identified Discrete Response Models," arXiv:2310.02414 [econ.EM] `https://arxiv.org/abs/2310.02414`.

Kim, J., and D. Pollard (1990): "Cube Root Asymptotics," *Annals of Statistics*, 18(1), 191–219.

Kitagawa, T., and A. Tetenov (2018): "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 86(2), 591–616.

Komarova, T. (2013): "Binary choice models with discrete regressors: Identification and misspecification," *Journal of Econometrics*, 177(1), 14–33.

Lee, S. M. S., and M. C. Pun (2006): "On m out of n Bootstrapping for Nonstandard M-Estimation With Nuisance Parameters," *Journal of the American Statistical Association*, 101(475), 1185–1197.

Manski, C. F. (1975): "Maximum score estimation of the stochastic utility model of choice," *Journal of Econometrics*, 3(3), 205–228.

——— (1985): "Semiparametric analysis of discrete response. Asymptotic properties of the maximum score estimator," *Journal of Econometrics*, 27(3), 313–333.

——— (1988): "Identification of binary response models," *Journal of the American Statistical Association*, 83(403), 729–738.

——— (2009): "The 2009 Lawrence R. Klein Lecture: Diversified Treatment under Ambiguity," *International Economic Review*, 50(4), 1013–1041.

Manski, C. F., and E. Tamer (2002): "Inference on regressions with interval data on a regressor or outcome," *Econometrica*, 70(2), 519–546.

Manski, C. F., and T. S. Thompson (1986): "Operational characteristics of maximum score estimation," *Journal of Econometrics*, 32(1), 85–108.

Patra, R. K., E. Seijo, and B. Sen (2015): "A Consistent Bootstrap Procedure for the Maximum Score Estimator," arXiv:1105.1976.

Pinkse, C. (1993): "On the computation of semiparametric estimates in limited dependent variable models," *Journal of Econometrics*, 58(1), 185–205.

Rosen, A. M., and T. Ura (2025): "Finite Sample Inference for the Maximum Score Estimand," *Review of Economic Studies*, forthcoming, https://academic.oup.com/restud/advance-article-pdf/doi/10.1093/restud/rdaf001/61372796/rdaf001.pdf.

Seo, M. H., and T. Otsu (2018): "Local M-estimation with Discontinuous Criterion for Dependent and Limited Observations," *Annals of Statistics*, 46(1), 344–369.

Walker, C. D. (2024): "A Bayesian Perspective on the Maximum Score Problem," arXiv:2410.17153 [econ.EM], https://arxiv.org/abs/2410.17153.