CAE Working Paper #05-10

Partial Identification of Probability Distributions with Misclassified Data

by

Francesca Molinari

April 2005

.

Partial Identification of Probability Distributions with Misclassified Data[*]

Francesca Molinari[†]
Cornell University[‡]

Abstract

This paper addresses the problem of data errors in discrete variables. When data errors occur, the observed variable is a misclassified version of the variable of interest, whose distribution is not identified. Inferential problems caused by data errors have been conceptualized through *convolution* and *mixture models*. This paper introduces the *direct misclassification approach*. The approach is based on the observation that in the presence of classification errors, the relation between the distribution of the "true" but unobservable variable and its misclassified representation is given by a linear system of simultaneous equations, in which the coefficient matrix is the matrix of misclassification probabilities. Formalizing the problem in these terms allows one to incorporate any prior information – e.g., validation studies, economic theory, social and cognitive psychology – into the analysis through sets of restrictions on the matrix of misclassification probabilities. Such information can have strong identifying power; the direct misclassification approach fully exploits it to derive identification regions for any real functional of the distribution of interest. A method for estimating the identification regions and construct their confidence sets is given, and illustrated with an empirical analysis of the distribution of pension plan types using data from the Health and Retirement Study.

Keywords: Misclassification; Partial Identification; Direct Misclassification Approach.

JEL Classification: C10, C13, C14, J26.

[*] First Draft: November 2002. This version: April 2005.

[†] I am grateful to Tim Conley, Joel Horowitz, Rosa Matzkin, and especially Chuck Manski for helpful comments and suggestions. I have benefitted from discussions with Talia Bar, Gadi Barlevy, Levon Barseghyan, Larry Blume, Riccardo DiCecio, Maria Goltsman, Ani Guerdjikova, George Jakubson, Nick Kiefer, Rasmus Lentz, Guido Menzio, Bruce Meyer, Marcin Peski, Jim Sullivan, Chris Taber, Elie Tamer, Tymon Tatur and Tim Vogelsang, and from the comments of seminar participants at Boston College, Chicago GSB, Cornell, Duke, Georgetown, Pittsburgh, Penn, Penn State, Princeton, Purdue, Toronto, UCLA, UCL, Virginia, and at the 2003 Southern Economic Association Meetings. All remaining errors are my own. Research support from Northwestern University Dissertation Year Fellowship and the Center for Analytic Economics at Cornell University is gratefully acknowledged.

[‡] Department of Economics, Cornell University, 492 Uris Hall, Ithaca NY 14853-7601.

1 Introduction

Error-ridden data constitute a significant problem in nearly all fields of science. There are many possible sources of data errors. Examples include use of inexact measures because of high costs or infeasibility of exact evaluation, tendency of study subjects to underreport socially undesirable behaviors and attitudes, and overreport socially desirable ones, or imperfect recall (or lack of knowledge) by study subjects. When data errors are present, often the sampling process does not identify the probability distribution of interest, and inference is impaired.

This paper addresses the problem of data errors in discrete variables. Interest in the question emerges from the observation that much of the empirical work in economics and related fields is based on the analysis of survey data. The reliability of these data is well documented to be less than perfect (see for example Bound, Brown, and Mathiowetz (2001)). Although survey questions may gather information on variables that are conceptualized as continuous (e.g.: age, earnings, etc.), a considerable part of the collected data is in the form of variables taking values in finite sets. Examples include educational attainment, language proficiency, workers' union status, employment status, health conditions and health/functional status.

When data errors occur in variables of this type, it is natural to think about the problem in terms of classification errors (see for example Bross (1954) and Aigner (1973)). An example may clarify this point. Suppose that an analyst is interested in learning the distribution of pension plan types in the American population. Three types are possible: defined benefit, defined contribution, and plans incorporating features of both. Suppose that the analyst has data from a nationally representative survey which queried a random sample of American households about their pension plans' characteristics. Validation studies document that a significant fraction of the reported plan types differ from the truth; for example, some people who truly have a defined benefit plan are erroneously classified as having a defined contribution plan (Gustman and Steinmeier (2001)).

To formalize the problem, suppose that each member l of a population L is characterized by the vector $(w_l, x_l) \in X \times X$, where X is a discrete set, not necessarily ordered, denoted by $X \equiv \{1, 2, \ldots, J\}$, $2 \leq J < \infty$. Let a sampling process draw persons at random from L. Suppose that the analyst is interested in learning features of the distribution $P(x)$ from the available data. However, she does not observe realizations of x, but observes realizations of w, which can either be equal or differ from the realizations of x. In the above example, x would denote the true pension plan type, and w the type reported in the survey.

Much of the existing literature on drawing inference in presence of error-ridden data has conceptualized the problem using either *convolution models* or *mixture models*. In the case of convolution models, a latent variable $v \in V$ is introduced, and w is assumed to measure x with chronic (i.e., affecting each observation) "errors-in-variables:" $w = x + v$. Researchers using convolution models commonly assume that the latent variable v is statistically independent from x or uncorrelated with x, and has mean zero (see, e.g., Klepper and Leamer (1984)).

In the case of mixture models, latent variables $v \in V$ and $z \in \{0,1\}$ are introduced, and w is viewed as a contaminated version of x, generated by the mixture $w = z \cdot x + (1-z) \cdot v$. In this model, the unobservable binary variable z denotes whether x or v is observed, and realizations of w with $z = 1$ are said to be error free. Researchers using mixture models commonly assume that the error probability $\Pr(z = 0)$ is known, or at least that it can be bounded non-trivially from above (see, e.g., Horowitz and Manski (1995)).

When a variable with finite support is imperfectly classified, it is widely recognized that the assumption, typical in convolution models, of independence between measurement error and true variable cannot hold (see for example Bound et al. (2001), p. 3735). Moreover, compelling evidence from validation studies suggests that errors in the data are occasional rather than "chronic:" a significant part of the observed data are error free. Mixture models seem therefore more suited for the analysis of such data. However, often the researcher has prior information on the nature of the misclassification pattern that has transformed x into w. This information may aid in identification, but cannot easily be exploited through a mixture model.

In this paper I propose an alternative framework, which I call the *direct misclassification approach*, to draw inference on the distribution of discrete variables subject to classification errors. The approach does not rely on the introduction of latent variables, but is based on the observation that in the presence of misclassification, the relation between the observable distribution of w and the unobservable distribution of x is given by

$$\begin{bmatrix} \Pr(w=1) \\ \vdots \\ \Pr(w=J) \end{bmatrix} = \begin{bmatrix} \Pr(w=1|\, x=1) & \dots & \Pr(w=1|\, x=J) \\ \vdots & \ddots & \vdots \\ \Pr(w=J|\, x=1) & \dots & \Pr(w=J|\, x=J) \end{bmatrix} \begin{bmatrix} \Pr(x=1) \\ \vdots \\ \Pr(x=J) \end{bmatrix}. \tag{1.1}$$

In all that follows I will denote by $\Pi^\star$ the matrix of elements $\{\Pr(w=i|\, x=j)\}_{i,j \in X}$ which appears on the right hand side of the above equation. For $i \neq j$, $\Pr(w=i|\, x=j)$ is generally referred to as "misclassification probability." Equation (1.1) is a simple formalism, and does not have content per se. However, it becomes potentially informative when combined with assumptions on the matrix of misclassification probabilities $\Pi^\star$; such assumptions generate a *misclassification model.*

The method that I introduce allows one to draw inference on $P(x)$ and on any real functional of this distribution using equation (1.1) directly, when restrictions on the elements of $\Pi^\star$ are imposed. Due to the classification errors, the identification of the probability distribution $P(x)$ is partial, and the inference on any of its real functionals is in the form of *identification regions*, that is, sets collecting the feasible values of such functionals. I show that these regions are "sharp," in the sense that they exhaust all the available information, given the sampling process and the maintained assumptions. Manski (2003) gives an overview of the literature on partial identification; for other work see e.g. Hotz, Mullin, and Sanders (1997) and Blundell, Gosling, Ichimura, and Meghir (2003).

The restrictions imposed on $\Pi^\star$ can have several origins, including validation studies, economic theory, cognitive and social psychology, or information on the circumstances under which the data have been collected. In this paper I study their identifying power in general. I then consider a few specific examples. As a starting point, I assume that the researcher has a known lower bound on the probability that the realizations of w and x coincide, i.e., $\Pr(w = x) \geq 1 - \lambda$, or, strengthening this assumption, that the researcher has a known lower bound on the probability of correct report for each value that x can take, i.e., $\Pr(w = j|\, x = j) \geq 1 - \lambda$, $\forall j \in X$. This information is often provided by validation studies or knowledge of the circumstances under which the data have been collected.[1] In this paper it is regarded as "base-case" information, and the identification regions derived under these assumptions constitute the baseline of the analysis. Then, I consider the case of "constant probability of correct report," and the case of "monotonicity in correct reporting." I show that these assumptions can have identifying power when maintained alone, as well as when imposed jointly with the base case assumptions.

The assumption of constant probability of correct report is motivated by the findings of validation studies. For specific survey inquiries, these studies suggest that the probability of correct report, for at least a subset of the values that x can take, is constant (formally, $\Pr(w = j|\, x = j) = \pi^\star$, $\forall j \in \tilde{X} \subseteq X$. In all that follows, I will denote by $\tilde{X} \subseteq X$ the subset of values that x can take, for which a given restriction holds). For example, in the context of self reports of employment status, Poterba and Summers' (1995) analysis suggests that there is approximately the same probability of correct report for people who are employed and for those who are not in the labor force, but a much lower probability of correct report for people who are unemployed.

The assumption of monotonicity in correct reporting is motivated by social psychology, which suggests that when survey respondents are asked questions relative to socially and personally sensitive topics, they tend to underreport socially undesirable behaviors and attitudes, and overreport socially desirable ones. This suggestion is supported by validation studies, which often document, within a given survey inquiry, that the probability of correct report of a certain alternative is greater or equal than the probability of correct report of a less socially desirable alternative (formally, $\Pr(w = j|\, x = j) \geq \Pr(w = j+1|\, x = j+1)$, $\forall j \in \tilde{X} \subset X$, where a higher value of j denotes a decrease in social desirability of an alternative). This is the case for example when survey respondents are asked about their participation in welfare programs, and j indicates non

[1] Availability of a lower bound on the error probability is a commonplace assumption in the statistic literature on robust estimation, which makes use of *mixture models*. For example, Hampel (1974) and Hampel et al. (1986) state that "the proportion of gross errors in data, depending on circumstances, is normally between 0.1% and 10% with several percent being the rule rather than the exception" (p. 387 and p. 28, respectively).

participation, while $j+1$ indicates participation (Bound et al. (2001) present a survey of validation studies on transfer program recipiency).

The proposed method allows the researcher to easily incorporate these assumptions, and in general any restriction on the misclassification pattern, into the analysis. The method is easy to implement, and often computationally tractable. Despite the fact that the results of validation studies on discrete variables are often presented in the form of matrices of misclassification probabilities (see, e.g., Bound et al. (2001)), and the appeal of the simple formalization given by the misclassification models, there appear to be no precedents to the direct use of equation (1.1) to deal with the identification problems caused by classification errors.

However, there are precedents to the use of specific restrictions on misclassification probabilities. Aigner (1973), Klepper (1988), and Bollinger (1996) imposed different sets of assumptions on the probabilities of misclassifying a dichotomous variable x, and derived sharp nonparametric bounds on the mean regression $E\left(y|\,x\right)$. Their approach is close in spirit to the one in this paper, but their methods are designed exclusively for binary variables, and for the case in which specific assumptions hold. On the other hand, most of the related literature (e.g.: Card (1996), Hausman, Abrevaya, and Scott-Morton (1998), Abrevaya and Hausman (1999), Lewbel (2000), Dustmann and van Soest (2000), Kane, Rouse, and Staiger (1999), Ramalho (2002)) proposes methods imposing restrictions on misclassification probabilities to achieve parametric or semiparametric identification of the quantities of interest (i.e., features of $P\left(y|\,x\right)$, or, less often, $P\left(x\right)$).[2] As such, these methods are subject to criticisms against possible misspecifications; moreover, while the assumptions employed might hold in some data sets, there might be other data sets for which they do not hold, and in that case the methods cannot be applied. Additionally, often these assumptions are maintained for technical reasons, and do not have an obvious interpretation.

Horowitz and Manski (1995) introduced fully nonparametric methods to draw inference on features of the distribution of a random variable x, when the sampling process is corrupted or contaminated. They adopted a mixture model, and showed that if the researcher has a (nontrivial) lower bound $1-\lambda$ on the probability that the realization of w is drawn from the distribution of x, informative bounds can be obtained on any parameter of the distribution $P\left(x\right)$ that respects stochastic dominance. Horowitz and Manski (1995) showed that these bounds are sharp, in the sense that they exhaust all the available information, given the sampling process and the maintained assumptions. The assumptions they entertain imply the base case assumptions on $\Pi^{\star}$ introduced

[2] Specific restrictions include the following: Bross (1954), when introducing the misclassification problem for binary data, assumed that $\Pr\left(w=1|\,x=0\right)$ and $\Pr\left(w=0|\,x=1\right)$ are of the same order of magnitude. Usually with binary data it is assumed either that $\Pr\left(w=1|\,x=0\right)=\Pr\left(w=0|\,x=1\right)<\frac{1}{2}$ (e.g., Klepper (1988), Card (1996)), or that $\Pr\left(w=1|\,x=0\right)+\Pr\left(w=0|\,x=1\right)<1$ (e.g., Bollinger (1996), Hausman et al. (1998)). When $J>2$, it is assumed that other monotonicity restrictions between the elements of $\Pi^{\star}$ hold (e.g., Abrevaya and Hausman (1999), Dustmann and van Soest (2000)), or that specific types of misclassification do not occur (Gong et al. (1990)).

above, namely $\Pr(w = x) \geq 1 - \lambda$, and $\Pr(w = j | x = j) \geq 1 - \lambda$, $\forall j \in X$.[3] When only these assumptions are maintained, in terms of identification of the types of parameters considered by Horowitz and Manski, the method developed in this paper is equivalent to the one they proposed.

However, often different, and perhaps more, information is available to the applied researcher beyond that maintained by Horowitz and Manski (1995). This information can have strong identifying power, but cannot be easily used within a mixture model. The direct misclassification approach allows one to readily incorporate it into the analysis, and fully exploit its identifying power. The method does not rely on any specific set of assumptions, but can incorporate any prior information that the researcher might have on the misreporting pattern into the analysis and guarantees sharpness of the implied identification regions.

While in the paper I focus on a single misclassified variable x, the method easily extends to drawing inference on features of the distribution of x conditional on a perfectly observed covariate, or on the joint distribution of several misclassified variables, taking values in finite sets. Given an outcome variable of interest $y \in Y$, the approach also extends to drawing inference on features of the distribution $P(y | x)$ when x is subject to classification errors. Moreover, it can allow one to draw inference when the data are not only error-ridden, but also incomplete, a situation very common in practice. In fact, in presence of both misclassified and missing data, the matrix in equation (1.1) will simply become rectangular rather than square, with additional rows giving the probabilities of having missing data, conditional on the true values of x.

The paper is organized as follows. Section 2 introduces the method, describes connectedness properties of the identification regions, outlines how the identification regions can be estimated consistently, and proposes a procedure to calculate confidence sets for the identification regions. Section 3 studies the identifying power of a few specific assumptions, some of which have not been previously considered in the literature. Section 4 illustrates the estimation method with an application to data on the distribution of pension plans' characteristics in the American population. Section 5 discusses the extensions of the direct misclassification approach mentioned above, showing how it allows the researcher to draw inference on features of the joint distribution of two or more variables, when one is perfectly measured, but at least another is subject to classification error. It also illustrates how to extend the method to the case of jointly missing and error ridden data. Section 6 concludes. An analysis of the relationship between misclassification models, convolution models, and mixture models is provided in Appendix A. All of the mathematical details are in Appendix B.

[3] If the researcher has an upper bound λ on the error probability, and the sampling process is corrupted, the first assumption follows; if the sampling process is contaminated, the second assumption follows. These results will be rigorously proved in Appendix A.

2 The Direct Misclassification Approach

In all that follows, to keep the focus on identification, I treat identified quantities as population parameters, and I assume that $\Pr(w = j) > 0 \ \forall \ j \in X$. A method to consistently estimate the identification regions and construct their confidence sets are provided at the end of this section.

Let $\mathbf{P}^w$ denote the column vector $\left[P_j^w, j \in X\right] \equiv [\Pr(w = j), j \in X]$, $\mathbf{P}^x$ the column vector $[\Pr(x = j), j \in X]$, and $\Pi^\star$ the stochastic matrix which, through equation (1.1), generates the misclassification of x into w. Denote the elements of $\Pi^\star$ by $\pi_{ij}^\star \equiv \{\Pr(w = i | x = j)\}$, $i, j \in X$, and the columns of $\Pi^\star$ by $\boldsymbol{\pi}_j^\star$. Let Ψ_X denote the space of all probability distributions on X, and define analogously $\Psi_{X \times W}$; let $\Re$ denote the real line. Let $\tau : \Psi_X \to \Re$ be a real functional of $P(x)$, denoted $\tau[\mathbf{P}^x]$, with analogous definitions for functionals of the joint distribution of (w, x). A particularly simple functional of $P(x)$ is $\tau[\mathbf{P}^x] = E[1(x = j)] = \Pr(x = j)$, $j \in X$. For any given matrix of functionals of interest Θ, let $H[\Theta]$ denote its identification region.

Given this notation, we can rewrite equation (1.1) as

$$\mathbf{P}^w = \Pi^\star \cdot \mathbf{P}^x. \tag{2.1}$$

The direct misclassification approach starts from the observation that $\Pr(x = j)$, $j \in X$, enters each of the J equations in system (1.1). Hence, each one of these equations can, potentially, imply restrictions on $\Pr(x = j)$, and therefore on $\mathbf{P}^x$ and $\tau[\mathbf{P}^x]$. The extent to which this will be the case crucially depends on what assumptions are imposed on the misreporting pattern.

The approach is quite intuitive. If $\Pi^\star$ were known, and of full rank, we would be able to solve the system of linear equations in (2.1) and uniquely identify $\mathbf{P}^x$, and therefore $\tau[\mathbf{P}^x]$. In practice, the misclassification probabilities $\pi_{ij}^\star$, $i, j \in X$, are known only to belong to a set $H[\Pi^\star]$, defined below. This set accounts both for the restrictions coming from probability theory, as well as for the restrictions on the misreporting pattern coming from validation studies, social and cognitive psychology, economic theory, etc. Denote the elements of $H[\Pi^\star]$ by $\Pi \equiv \{\pi_{ij}\}_{i,j \in X}$, and the columns of this matrix by $\boldsymbol{\pi}_j$, $j \in X$. When $H[\Pi^\star]$ is not a singleton, $\mathbf{P}^x$ is not identified and $\tau[\mathbf{P}^x]$ need not be identified, but only known, respectively, to lie in the identification regions $H[\mathbf{P}^x]$ and $H\{\tau[\mathbf{P}^x]\}$.

The identification region $H[\mathbf{P}^x]$ is defined as the set of column vectors $\mathbf{p}^x = [p_k^x, k \in X]$, such that, given $\Pi \in H[\Pi^\star]$, $\mathbf{p}^x$ solves system (2.1):

$$H[\mathbf{P}^x] = \{\mathbf{p}^x : \mathbf{P}^w = \Pi \cdot \mathbf{p}^x, \ \Pi \in H[\Pi^\star]\}. \tag{2.2}$$

In the next Subsection, $H[\Pi^\star]$ will be formally defined, and characterized in a way such that $\forall$ $\Pi \in H[\Pi^\star]$, $p_k^x \geq 0$, $\forall \ k \in X$, and $\sum_{k=1}^J p_k^x = 1$.

Throughout this paper, the notation $\mathbf{p}^x$ will be reserved to elements of $H[\mathbf{P}^x]$, and the notation p_k^x to the $k-$th component of a vector $\mathbf{p}^x$. Hence, p_k^x and $\mathbf{p}^x$ represent, respectively, feasible values of

$\Pr(x=k)$, $k \in X$, and $[\Pr(x=j), j \in X]$, given $\Pi \in H[\Pi^\star]$ and equation (2.1). By construction

$$\begin{aligned}\mathbf{p}^x &\equiv \mathbf{p}^x(\Pi; \mathbf{P}^w),\\ p_k^x &= p_k^x(\Pi; \mathbf{P}^w),\ k \in X.\end{aligned}$$

For ease of notation, I omit the arguments of p_k^x and $\mathbf{p}^x$. The identification region $H\{\tau[\mathbf{P}^x]\}$ is then defined as:

$$H\{\tau[\mathbf{P}^x]\} = \{\tau[\mathbf{p}^x] : \mathbf{p}^x \in H[\mathbf{P}^x]\}. \tag{2.3}$$

The set $H[\Pi^\star]$ is of central importance for the identification of $\mathbf{P}^x$ and $\tau[\mathbf{P}^x]$, as the identification regions of these functionals are defined on the basis of $H[\Pi^\star]$. I denote by $H^P[\Pi^\star]$ the set of matrices that satisfy the probabilistic constraints and by $H^E[\Pi^\star]$ the set of matrices satisfying the constraints coming from validation studies and theories developed in the social sciences. Hence,

$$H[\Pi^\star] = H^P[\Pi^\star] \cap H^E[\Pi^\star]$$

In what follows, I will describe the geometry of $H[\Pi^\star]$, and in particular its connectedness properties. Interest in connectedness arises from the fact that the continuous image of a connected set is connected. This implies that if $H[\Pi^\star]$ is connected and $\mathbf{p}^x$ is a continuous function of Π, $H[\mathbf{P}^x]$ is connected as well, and so is $H\{\tau[\mathbf{P}^x]\}$ if $\tau(\cdot)$ is a continuous functional. Conversely, if $H[\Pi^\star]$ is not connected or if the functionals are not continuous, $H[\mathbf{P}^x]$ and $H\{\tau[\mathbf{P}^x]\}$ need not necessarily be connected. This has implications for the estimation of the identification regions. Consider for example the case that interest centers on a real valued functional $\tau[\mathbf{P}^x]$. When $H\{\tau[\mathbf{P}^x]\}$ is a connected set, it is given by the entire interval between its smallest and its largest points. Hence by estimating these two points one obtains an estimate of the entire identification region. When $H\{\tau[\mathbf{P}^x]\}$ is disconnected, parts of the interval between the smallest and the largest points are not feasible, and therefore are not elements of the identification region. Section 2.2 introduces a method to estimate $H\{\tau[\mathbf{P}^x]\}$ when this is the case.

A relevant example of a case in which $\mathbf{p}^x$ is a continuous function of Π is obtained when each matrix $\Pi \in H[\Pi^\star]$ is of full rank. In this case, for each $\Pi \in H[\Pi^\star]$, one can solve the linear system in (2.1), obtaining $\mathbf{p}^x = \Pi^{-1} \cdot \mathbf{P}^w$. It is a well known result in matrix algebra that the inverse of a nonsingular matrix is continuous in the elements of the matrix (see, e.g., Campbell and Meyer (1991) Ch. 10). A very simple condition ensuring that each matrix $\Pi \in H[\Pi^\star]$ is of full rank is assuming that the probability of correct report is greater than $\frac{1}{2}$ for each of the values that x can take.[4] Validation studies suggest that this requirement is often satisfied in practice.[5]

[4] If $\pi_{jj} > \frac{1}{2}, \forall j \in X$, $\forall\ \Pi \in H[\Pi^\star]$, Π^T is strictly diagonally dominant, and hence Π is nonsingular. An $n \times n$ matrix $A = \{a_{ij}\}$ is said to be strictly diagonally dominant if, for $i = 1, 2, \ldots, n$, $|a_{ii}| > \sum_{j=1(j\neq i)}^{n} |a_{ij}|$. A proof of the fact that if A is strictly diagonally dominant, then A is nonsingular, can be found in Horn and Johnson (1999), Theorem 6.1.10.

[5] Among others, this is the case in the context of workers' union status (see, e.g., Card (1996)), transfer program

2.1 The Set $H[\Pi^{\star}]$ and its Geometry

We start by characterizing the set $H^P[\Pi^{\star}]$ and its geometry. Probability theory requires that $\sum_{i=1}^{J}\pi_{ij} = 1$, $\forall j \in X$, that $\pi_{ij} \geq 0$, $\forall i, j \in X$, and that, given $\mathbf{P}^w$, equation (2.1), and Π, the implied $\mathbf{p}^x$ gives a valid probability measure. Denote by $H^P[\Pi^{\star}]$ the set of Πs that satisfy these probabilistic requirements, so that, throughout the entire paper,

$$H^P[\Pi^{\star}] \equiv \left\{\Pi : \begin{pmatrix} \pi_{ij} \geq 0,\ \forall i, j \in X, & \sum_{i=1}^{J}\pi_{ij} = 1,\ \forall j \in X, \\ p_h^x \geq 0\ \forall h \in X, & \sum_{h=1}^{J} p_h^x = 1 \end{pmatrix}\right\}. \qquad (2.4)$$

Notice that the set $H^P[\Pi^{\star}]$ can be defined alternatively using the notions of $(J-1)-$ dimensional simplex and convex hull of a set of vectors. We will use the following definitions:

Definition 1 *The $(J-1)-$dimensional simplex is the set* $\Delta_{J-1} \equiv \left\{\boldsymbol{\delta} \in \Re_+^J : \delta_1 + \delta_2 + \ldots + \delta_J = 1\right\}$.

Definition 2 *The convex hull of a finite subset $\{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_J\}$ of $\Re^J$, denoted conv $\{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_J\}$, consists of all the vectors of the form $\alpha_1\boldsymbol{\xi}_1 + \alpha_2\boldsymbol{\xi}_2 + \ldots + \alpha_J\boldsymbol{\xi}_J$ with $\alpha_i \geq 0\ \forall\ i = 1, \ldots, J$ and $\sum_{i=1}^{J}\alpha_i = 1$. (Rockafellar (1970), Corollary 2.3.1.)*

By definition, $\mathbf{P}^w \in \Delta_{J-1}$. We can now rewrite the set $H^P[\Pi^{\star}]$ as

$$H^P[\Pi^{\star}] \equiv \left\{\Pi : \boldsymbol{\pi}_j \in \Delta_{J-1} \text{ and } p_j^x \geq 0\ \forall j \in X, \text{ and } \mathbf{P}^w \in conv\left\{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_J\right\}\right\}. \qquad (2.5)$$

In words, a matrix Π is an element of $H^P[\Pi^{\star}]$ if its columns are probability mass functions, the implied $\mathbf{p}^x$ is nonnegative, and the vector $\mathbf{P}^w$ can be expressed as a convex combination of the columns of Π.

To describe the geometry of $H^P[\Pi^{\star}]$ we need to introduce another definition:

Definition 3 *A subset Γ of $\Re^n$ is star convex with respect to $\boldsymbol{\gamma}_0 \in \Gamma$ if for each $\boldsymbol{\gamma} \in \Gamma$ the line segment joining $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_0$ lies in Γ. (Munkres (1991), p. 330.)*

As a remark, a star convex set is always pathwise-connected, which in turn is always connected. Given a set of matrices $H^P[\Pi^{\star}] \subset \Re^{J\times J}$, I will define the line segment between two matrices $\Pi^1, \Pi^2 \in H^P[\Pi^{\star}]$ as

$$\Pi^{\alpha} = \alpha\Pi^1 + (1-\alpha)\Pi^2,\ \alpha \in [0,1],$$

and I will say that the set $H^P[\Pi^{\star}]$ is convex if given any two matrices $\Pi^1, \Pi^2 \in H^P[\Pi^{\star}]$, $\Pi^{\alpha} \in H^P[\Pi^{\star}]$ for all $\alpha \in (0,1)$. Given these preliminaries, let $\tilde{\Pi}$ be a matrix with each column identical to $\mathbf{P}^w$, and notice that $\tilde{\Pi}$ is trivially in $H^P[\Pi^{\star}]$. We are now ready to state a result describing the connectedness of the set $H^P[\Pi^{\star}]$.

recipiency (see, e.g., Moore, Marquis, and Bogen (1996)), employment status (see, e.g., Poterba and Summers (1995)), and 1- and 3-digit level classification of industry and occupation (see, e.g., Mellow and Sider (1983)).

Proposition 1 *The set $H^P[\Pi^\star]$ is star convex with respect to $\tilde{\Pi}$. However, it is not star convex with respect to any other of its elements.* $\square$

The result in Proposition 1 implies that the set $H^P[\Pi^\star]$ is not convex, because a convex set is star convex with respect to each of its elements. The set $H^P[\Pi^\star]$ is illustrated in Example 1 and in the first panel of Figure 1.

Example 1 *Suppose that x and w are binary, i.e. that $J = 2$, and let $P_1^w = 0.3$. Then the matrix Π is determined by its two diagonal elements, π_{11} and π_{22}, and*

$$p_1^x = \frac{P_1^w - (1 - \pi_{22})}{\pi_{11} - (1 - \pi_{22})}.$$

It is easy to verify that

$$H^P[\Pi^\star] = \{\pi_{11}, \pi_{22} : (\pi_{11} \in [0, P_1^w], \pi_{22} \in [0, 1 - P_1^w]) \cup (\pi_{11} \in [P_1^w, 1], \pi_{22} \in [1 - P_1^w, 1])\}.$$

This set is plotted in the first panel of Figure 1, and its star convexity is apparent.

Let us now turn to the set of matrices, denoted $H^E[\Pi^\star]$, that satisfy the restrictions on the misreporting pattern coming from prior information. Then if, for example, validation studies suggest a uniform lower bound on the probability of correct report for each $j \in X$, we will have

$$H^E[\Pi^\star] = \{\Pi : \pi_{jj} \geq 1 - \lambda \ \forall j \in X\}.$$

If social psychology suggests that individuals, when answering about the frequency with which they engage in a certain socially desirable activity, either provide correct reports or over-report, we will have

$$H^E[\Pi^\star] = \{\Pi : \pi_{ij} = 0 \ \forall \ i < j \in X\}.$$

Of course, plenty of other restrictions are possible.

Let us now return to Proposition 1, and analyze the insight that it provides. Since $H^P[\Pi^\star]$ is connected, but not convex, when we take its intersection with the set $H^E[\Pi^\star]$ we obtain a set $H[\Pi^\star]$ that might be disconnected, connected, or convex, depending on how $H^E[\Pi^\star]$ slices $H^P[\Pi^\star]$. Below I provide three examples of sets $H^E[\Pi^\star]$, that will further be analyzed in Section 3. Each of these sets is trivially convex, as it is linear in Π, but its intersection with $H^P[\Pi^\star]$ generates sets $H[\Pi^\star]$ that can be disconnected, connected, and convex. These examples are illustrated in the six panels of Figure 1.

Example 2 ***Constant Probability of Correct Report.***
Let $H^E[\Pi^\star] = \{\Pi : \pi_{jj} = \pi,\ \forall j \in X\}$. Suppose that x and w are binary, i.e. that $J = 2$. Then

$$H[\Pi^\star] = \begin{cases} \{\pi : \pi \in [0, P_1^w] \cup [1 - P_1^w, 1]\} & \textit{if } P_1^w < \frac{1}{2}, \\ \{\pi : \pi \in [0, 1 - P_1^w] \cup [P_1^w, 1]\} & \textit{if } P_1^w > \frac{1}{2}, \\ \{\pi : \pi \in [0, 1]\} & \textit{if } P_1^w = \frac{1}{2}. \end{cases}$$

Hence, if $P_1^w \neq \frac{1}{2}$, $H[\Pi^\star]$ is disconnected. This set is plotted in the second panel of Figure 1, and the fact that it is disconnected is apparent. Moreover, it is apparent that the set $H[\Pi^\star]$ will remain disconnected, if $P_1^w \neq \frac{1}{2}$, even if the assumption of constant probability of correct report is weakened to requiring that $\pi_{22} = \pi_{11} + \varepsilon$, as long as $|\varepsilon| < |1 - 2P_1^w|$ (and ε is such that $\pi_{22} \in [0, 1]$).

Example 3 ***Monotonicity in Correct Reporting.***
Let $H^E[\Pi^\star] = \left\{\Pi : \pi_{jj} \geq \pi_{(j+1)(j+1)},\ \forall j \in X\right\}$. Suppose that x and w are binary, i.e. that $J = 2$, so that the monotonicity assumption simplifies to $\pi_{11} \geq \pi_{22}$. Then if $P_1^w < \frac{1}{2}$,

$$H[\Pi^\star] = \{\pi_{11}, \pi_{22} : (\pi_{11} \in [0, P_1^w], \pi_{22} \in \{[0, \pi_{11}]\}) \cup (\pi_{11} \in [1 - P_1^w, 1], \pi_{22} \in [1 - P_1^w, \pi_{11}])\}$$

If $P_1^w \geq \frac{1}{2}$,

$$H[\Pi^\star] = \{\pi_{11}, \pi_{22} : (\pi_{11} \in [0, P_1^w], \pi_{22} \in [0, \min(1 - P_1^w, \pi_{11})]) \cup (\pi_{11} \in [P_1^w, 1], \pi_{22} \in [1 - P_1^w, \pi_{11}])\}$$

Hence, if $P_1^w < \frac{1}{2}$, $H[\Pi^\star]$ is disconnected, but otherwise it is connected. This set is plotted in the third panel of Figure 1. The fact that it is disconnected is apparent given the choice of $P_1^w = 0.3$. To see why the set can be connected, the fourth panel of Figure 1 plots the set $H[\Pi^\star]$ that would be obtained if the monotonicity assumption was $\pi_{11} \leq \pi_{22}$ (in the binary case, reversing the sign of the monotonicity assumption has an effect similar to maintaining $\pi_{11} \geq \pi_{22}$ but having $P_1^w > \frac{1}{2}$).

Example 4 ***Lower Bound on the Probability of Correct Report.***
Let $H^E[\Pi^\star] = \{\Pi : \pi_{jj} \geq 1 - \lambda, \forall j \in X\}$. Suppose that x and w are binary, i.e. that $J = 2$. Then if $1 > \lambda > \max\{P_1^w, 1 - P_1^w\}$,

$$H[\Pi^\star] = \{\pi_{11}, \pi_{22} : (\pi_{11} \in [1 - \lambda, P_1^w], \pi_{22} \in [1 - \lambda, 1 - P_1^w]) \cup (\pi_{11} \in [P_1^w, 1], \pi_{22} \in [1 - P_1^w, 1])\}.$$

This set is connected through the point $\pi_{11} = P_1^w$, $\pi_{22} = 1 - P_1^w$, and is plotted in the fifth panel of Figure 1 for $P_1^w = 0.3$ and $\lambda = 0.8$.
If $\max\{P_1^w, 1 - P_1^w\} > \lambda$, then

$$H[\Pi^\star] = \{\pi_{11}, \pi_{22} : \pi_{11} \in [\max\{1 - \lambda, P_1^w\}, 1], \pi_{22} \in [\max\{1 - \lambda, 1 - P_1^w\}, 1]\},$$

and $H[\Pi^\star]$ is convex. This set is plotted in the sixth panel of Figure 1, and the fact that it is convex is apparent given the choice of $P_1^w = 0.3$ and $\lambda = 0.2$.

2.2 Consistent Estimation of the Identification Regions

Suppose first that the researcher is simply interested in the extreme points of the identification region of a functional of $\mathbf{P}^x$, say for example $\tau\left[\mathbf{P}^x\right] = \Pr\left(x = j\right), j \in X$, and that the matrix Π is of full rank for any $\Pi \in H\left[\Pi^\star\right]$. Then these points can be calculated and consistently estimated by solving nonlinear optimization problems subject to linear and nonlinear constraints. In particular, let $\mathbf{p}^x = \Pi^{-1} \cdot \mathbf{P}^w$, $\Pi \in H\left[\Pi^\star\right]$. Then the smallest and the largest points in $H\left[\Pr\left(x = j\right)\right], j \in X$, can be calculated as

$$p_j^{x,L} = \inf_{\Pi \in H[\Pi^\star]} p_j^x, \qquad p_j^{x,U} = \sup_{\Pi \in H[\Pi^\star]} p_j^x,$$

and similarly for any other real functional. These extreme points are continuous functions of $\mathbf{P}^w$. Suppose for simplicity that only $\mathbf{P}^w$ needs to be estimated, and that a random sample $\{w_i\}$, $i = 1, \ldots, N$ is available. Let $\mathbf{P}_N^w$ be the vector collecting the fraction of observations reporting $w = i$, $i = 1, \ldots, J$,

$$P_{i,N}^w = \frac{1}{N} \sum_{j=1}^N 1\left(w_j = i\right), \qquad i = 1, \ldots, J. \tag{2.6}$$

Then one can consistently estimate the above extreme points by replacing $\mathbf{P}^w$ with $\mathbf{P}_N^w$.

Suppose now that the researcher is interested in estimating the entire identification region. While the general identification approach proposed in Section 2.1 is valid for any set of restrictions on $\Pi^\star$, here I will focus on restrictions that satisfy certain regularity conditions, described in Assumptions C0 and C1 below, so that a simple estimator can be utilized.

We have seen in the previous section that the set $H\left[\Pi^\star\right]$ can be disconnected, connected or convex. These properties will be reflected in the shape of the identification regions of the functionals that we are interested in, namely $H\left[\mathbf{P}^x\right]$, $H\left\{\tau\left[\mathbf{P}^x\right]\right\}$ and $H\left\{\Theta\left[\mathbf{P}^x\right]\right\}$, for some vector of dimension k of functionals $\Theta : \Psi_X \to \Re^k$. Hence, it is important to have a method to calculate and consistently estimate the entire identification regions, that will be able to capture their possible disconnectedness and nonconvexities.

Manski and Tamer (2002) introduced methods to estimate the entire identification region of a vector of parameters of interest when the identification region cannot be expressed in closed form solution, but is given by all values of the vector that minimize a specified objective function. Here I introduce a related nonlinear programming estimator, using the same insight as in the linear programming estimator proposed by Honore and Tamer (2003) and further discussed by Honore and Lleras-Muney (2004). Observe that if we can calculate $H\left[\mathbf{P}^x\right]$, we can then calculate $H\left\{\tau\left[\mathbf{P}^x\right]\right\}$ and $H\left\{\Theta\left[\mathbf{P}^x\right]\right\}$ for any functionals $\tau\left(\cdot\right)$ and $\Theta\left(\cdot\right)$ (for example, the mean of x, its variance, the Gini coefficient, etc.); hence, we focus on the calculation of $H\left[\mathbf{P}^x\right]$.

The set $H\left[\mathbf{P}^x\right]$ consists of the vectors $\mathbf{p}^x \in \Delta_{J-1}$ for which the equations

$$\begin{cases} \mathbf{P}^w = \Pi \cdot \mathbf{p}^x, \\ \boldsymbol{\pi}_j \in \Delta_{J-1} \ \forall j \\ \Pi \in H^E\left[\Pi^\star\right], \end{cases} \tag{2.7}$$

have a solution for Π. In general, $H^E\left[\Pi^\star\right]$ can be written as

$$H^E\left[\Pi^\star\right] = \left\{ \begin{array}{c} \Pi : f_j\left(\Pi\right) \geq \mu_j, \ j = 1, \ldots, q_1, \ g_i\left(\Pi\right) \leq \mu_{q_1+i}, \ i = 1, \ldots, q_2, \\ h_k\left(\Pi\right) = \mu_{q_1+q_2+k}, \ k = 1, \ldots, q_3, \end{array} \right\}$$

where $q_1 + q_2 + q_3 = q$ is the number of constraints imposed, and for $j = 1, \ldots, q$, $0 \leq \mu_j \leq M$ is a non-negative bounded parameter, and f_j: $\Re^{J^2} \longrightarrow \Re$, g_i: $\Re^{J^2} \longrightarrow \Re$, and h_k: $\Re^{J^2} \longrightarrow \Re$, are functions taking as arguments the elements of the matrix Π.

To give a concrete example, if $X = \{1, 2, 3\}$ and

$$H^E\left[\Pi^\star\right] = \left\{\Pi : \pi_{jj} \geq 0.8 \ \forall j \in X, \ 0.125 \leq \pi_{12}\pi_{13} \leq 0.33, \ \pi_{11} = \pi_{22}\right\},$$

then $q_1 = 4$, $q_2 = 1$, $q_3 = 1$, $q = 6$, and

$$\begin{array}{rcll} f_j\left(\Pi\right) & = & \pi_{jj}, & \mu_j = 0.8, \ j = 1, 2, 3, \\ f_4\left(\Pi\right) & = & \pi_{12}\pi_{13}, & \mu_4 = 0.125, \\ g_1\left(\Pi\right) & = & \pi_{12}\pi_{13}, & \mu_5 = 0.33, \\ h_1\left(\Pi\right) & = & \pi_{11} - \pi_{22}, & \mu_6 = 0. \end{array}$$

The equations in (2.7) have the same structure as the constraints in a nonlinear programming problem. Hence one can check whether a particular vector $\boldsymbol{\xi} \in \Delta_{J-1}$ belongs to $H\left[\mathbf{P}^x\right]$ by checking if a nonlinear programming problem that has constraints given by (2.7) has a solution with a specific value for the objective function. Consider the nonlinear programming problem

$$\max_{\{\pi_{ij}\},\{v_k\}} \sum_k -v_k \tag{2.8}$$

subject to

$$\begin{cases} v_k \geq 0 \ \ \forall \ k, \\ \pi_{ij} \geq 0, \ i, j = 1, \ldots, J, \\ 1 - \sum_{i=1}^J \pi_{ij} = v_j, \ j = 1, \ldots, J, \\ \mathbf{P}^w - \Pi \cdot \boldsymbol{\xi} = \left[\begin{array}{ccc} v_{J+1} & \ldots & v_{2J} \end{array}\right]^T, \\ f_l\left(\Pi\right) - \mu_l + v_{2J+l} \geq 0, \ l = 1, \ldots, q_1, \\ \mu_{q_1+m} - g_m\left(\Pi\right) + v_{2J+q_1+m} \geq 0, \ l = 1, \ldots, q_2, \\ h_s\left(\Pi\right) - \mu_{q_1+q_2+s} + v_{2J+q_1+q_2+s} = 0, \ l = 1, \ldots, q_3. \end{cases} \tag{2.9}$$

We will consider restrictions determining the set $H^E[\Pi^\star]$ that satisfy the following conditions:

Assumption C0: For each $j = 1,\ldots,q_1$, $i = 1,\ldots,q_2$, and $k = 1,\ldots,q_3$, $f_j(\Pi)|_{\Pi=\mathbf{0}} = g_i(\Pi)|_{\Pi=\mathbf{0}} = h_k(\Pi)|_{\Pi=\mathbf{0}} = \mathbf{0}$ and $f_j(\Pi)$, $g_i(\Pi)$, and $h_k(\Pi)$ are continuous on $[0,1]^{J^2}$.

Let $\mathcal{P}\times\mathcal{V}$ denote the constraint set defined by (2.9). Then under Assumption C0, $\mathcal{P}\times\mathcal{V}$ is closed, as the functions defining it are continuous. It is also non-empty, as it contains the vector $\left[\boldsymbol{\pi}_1^0;\ldots;\boldsymbol{\pi}_J^0;\mathbf{v}^0\right]$, with $\pi_{ij}^0 = 0$ for $i,j = 1,\ldots,J$, $v_j^0 = 1$ for $j = 1,\ldots,J$, $v_{J+j}^0 = P_j^w$ for $j = 1,\ldots,J$, $v_{2J+l}^0 = \mu_l$, $l = 1,\ldots,q_1$, $v_{2J+q_1+m}^0 = 0$, $m = 1,\ldots,q_2$, and $v_{2J+q_1+q_2+s}^0 = \mu_{q_1+q_2+s}$, $s = 1,\ldots,q_3$. The objective function in (2.8) is continuous. Moreover, the set

$$\left\{[\boldsymbol{\pi}_1;\ldots;\boldsymbol{\pi}_J;\mathbf{v}] \in \mathcal{P}\times\mathcal{V} : \textstyle\sum_k -v_k \geq \sum_k -v_k^0\right\}$$

is bounded. Hence, by the Bolzano-Weierstrass theorem, the objective function in (2.8) achieves a maximum on (2.9). The optimal function will have value zero if and only if all $v_k = 0$, that is if a solution exists to (2.7). Hence, for given $\boldsymbol{\xi} \in \Delta_{J-1}$ one can check whether $\boldsymbol{\xi} \in H[\mathbf{P}^x]$ by solving the above nonlinear programming problem and checking whether $v_k = 0$ for all k.

The above method for calculating identification regions has a natural sample analog counterpart, and under some regularity conditions about the functions defining the set $H^E[\Pi^\star]$ and the sampling process, this estimator is consistent. In particular, we will maintain the following assumptions:

Assumption C1: For each $j = 1,\ldots,q_1$, $i = 1,\ldots,q_2$, and $k = 1,\ldots,q_3$, either (i) $f_j(\Pi)$, $g_i(\Pi)$ and $h_k(\Pi)$ are homogeneous functions of degree (respectively) $r_j, r_i, r_k \geq 1$, or (ii) $f_j(\Pi)$, $g_i(\Pi)$ and $h_k(\Pi)$ are multivariate polynomials in Π with non-negative coefficients. Additionally, $g_i(\Pi) \geq 0$ and $h_k(\Pi) \geq 0$ on $[0,1]^{J^2}$.

Assumption C2: (a) Let a random sample $\{w_i\}$, $i = 1,\ldots,N$ be available, and let $\mathbf{P}_N^w$ be defined as in (2.6). (b) If the set $H^E[\Pi^\star]$ contains constraints involving any parameters to be estimated, let these parameters enter the constraints additively. Without loss of generality, to simplify the notation, let the parameters to be estimated be μ_l, $l = 1,\ldots,\bar{q} \leq q$. (c) Suppose that a random sample of size $n = \frac{N}{\kappa}$ for some constant κ such that $0 < \kappa < \infty$ is available to estimate μ_l, $l = 1,\ldots,\bar{q}$, so that $\sqrt{N}\left(\mu_{l,n} - \mu_l\right) \xrightarrow{d} N\left(0, \kappa V_{\mu_l}\right)$. (d) Let μ_l satisfy $\mu_l > 0, l = 1,\ldots,\bar{q} \leq q$.

In Section 3 we will consider several examples of restrictions defining the set $H^E[\Pi^\star]$ that satisfy Assumptions C0-C1. For example, suppose that a validation study provides a lower bound on the probability of correct report for each type $j = 1,\ldots,J$, so that $H^E[\Pi^\star] = \left\{\Pi : \pi_{jj} \geq \mu_j,\ j \in X\right\}$. Then Assumptions C0-C1 are clearly satisfied. Moreover, if a validation (random) sample $\{\tilde{w}_i, \tilde{x}_i\}$, $i = 1,\ldots,n$ is available (with $n = \frac{N}{\kappa}$, $0 < \kappa < \infty$), Assumption C3 is satisfied, and $\mu_{j,n}$ can be obtained as:

$$\mu_{j,n} = \frac{\sum_{i=1}^n 1\left(\tilde{w}_i = j, \tilde{x}_i = j\right)}{\sum_{i=1}^n 1\left(\tilde{x}_i = j\right)}$$

Let $H_N^E[\Pi^\star]$ denote the set $H^E[\Pi^\star]$ obtained when μ_l is replaced by $\mu_{l,n}$, $l = 1, \ldots, q$, with the convention that $\mu_{l,n} = \mu_l$ for $l = \bar{q}+1, \ldots, q$. Define an objective function $Q_N(\boldsymbol{\xi})$ by

$$Q_N(\boldsymbol{\xi}) = \max_{\{\pi_{ij}\},\{v_k\}} \sum_k -v_k$$

subject to

$$\begin{cases} v_k \geq 0 \;\; \forall\, k \\ \pi_{ij} \geq 0 \;\; \forall\, i,j = 1,\ldots,J \\ 1 - \sum_{i=1}^{J} \pi_{ij} = v_j, \;\; j = 1,\ldots,J \\ \mathbf{P}_N^w - \Pi \cdot \boldsymbol{\xi} = \left[\begin{array}{ccc} v_{J+1} & \ldots & v_{2J} \end{array} \right]^T \\ f_l(\Pi) - \mu_{l,n} + v_{2J+l} \geq 0, \;\; l = 1,\ldots,q, \\ \mu_{(q_1+m),n} - g_m(\Pi) + v_{2J+q_1+m} \geq 0, \;\; l = 1,\ldots,q_2, \\ h_s(\Pi) - \mu_{(q_1+q_2+s),n} + v_{2J+q_1+q_2+s} = 0, \;\; l = 1,\ldots,q_3. \end{cases}$$

Let $Q(\boldsymbol{\xi})$ be defined similarly, using (2.8)-(2.9). Then we have the following consistency result:

Proposition 2 *Let Assumptions C0, C1 and C2 hold. Define the set*

$$H_N[\mathbf{P}^x] = \left\{ \mathbf{p}_N^x \in \Delta_{J-1} : Q_N(\mathbf{p}_N^x) \geq \sup_{\boldsymbol{\xi} \in \Delta_{J-1}} Q_N(\boldsymbol{\xi}) - \epsilon_N \right\}, \qquad (2.10)$$

where $\epsilon_N = N^{-\tau}$, $0 < \tau < \frac{1}{2}$. Then the set $H_N[\mathbf{P}^x]$ is a consistent estimator of $H[\mathbf{P}^x]$, in the sense that

$$\begin{aligned} \rho(H_N[\mathbf{P}^x], H[\mathbf{P}^x]) &\equiv \sup_{\mathbf{p}_N^x \in H_N[\mathbf{P}^x]} \inf_{\mathbf{p}^x \in H[\mathbf{P}^x]} \|\mathbf{p}_N^x - \mathbf{p}^x\| \to_p 0, \\ \rho(H[\mathbf{P}^x], H_N[\mathbf{P}^x]) &\equiv \sup_{\mathbf{p}^x \in H[\mathbf{P}^x]} \inf_{\mathbf{p}_N^x \in H_N[\mathbf{P}^x]} \|\mathbf{p}_N^x - \mathbf{p}^x\| \to_p 0. \end{aligned}$$

Proof. See Appendix B. ■

Most of the calculations and estimations of $H[\mathbf{P}^x]$ presented in this paper are performed using this nonlinear programming method.

2.3 Confidence Sets for the Identification Regions[6]

The problem of the construction of confidence intervals for partially identified parameters was addressed by Horowitz and Manski (1998, 2000). They considered the case in which the identification region of the parameter of interest is an interval whose lower and upper bounds can be estimated from sample data, and proposed confidence intervals that asymptotically cover the entire identification region with fixed probability. For the same class of problems, Imbens and Manski (2004)

[6] I am very grateful to Elie Tamer for suggestions that led to the construction of these confidence sets.

suggested shorter confidence intervals that uniformly cover the parameter of interest, rather than its identification region, with a prespecified probability. These approaches are not applicable to the problem studied here, because our identification regions are given by the set of values of the parameters of interest that solve a minimization problem, and do not have a closed form solution. The problem of construction of confidence sets for identification regions of parameters obtained as the solution of the minimization of a criterion function has recently been addressed by Chernozhukov, Hong, and Tamer (2004). They provided a method to construct confidence sets that cover the identification region with probability asymptotically equal to $(1-\alpha)$, and developed a new subsampling bootstrap method to implement this procedure. Here I consider a different procedure, and show that the coverage property of these confidence sets follow directly from well known results in the literature (e.g., Rao (1973), Cox and Hinkley (1974)). The counterpart of the simplicity of this approach is that the confidence sets may be conservative, in the sense that given a prespecified confidence coefficient $(1-\alpha)$, $0<\alpha<1$, the confidence sets will asymptotically cover the identification region with probability at least equal to $(1-\alpha)$.

The main insight for the construction of the confidence sets for $H[\mathbf{P}^x]$, denoted $C_N^{H[\mathbf{P}^x]}$, is given by observing that the only parameters to be estimated for obtaining $H_N[\mathbf{P}^x]$ in (2.10) are $P_{i,N}^w$, $i=1,\ldots,J-1$, and $\mu_{l,n}$, $l=1,\ldots,\bar{q}$. Let $\hat{\boldsymbol{\vartheta}}_N$ denote the $J-1+\bar{q}$ vector collecting these estimators. Under Assumption C2, $\hat{\boldsymbol{\vartheta}}_N$ is root-N consistent and asymptotically normal, and has a covariance matrix $(Var(\boldsymbol{\vartheta}))$ that can be consistently estimated from the data $(\widehat{Var}\left(\hat{\boldsymbol{\vartheta}}_N\right))$. Hence, if $c_{1-\alpha}$ denotes the $(1-\alpha)$ quantile of the $\chi^2_{(J-1+\bar{q})}$ distribution, we can construct a joint confidence ellipsoid for $\boldsymbol{\vartheta}\equiv\left[(P_i^w)_{i=1,\ldots,J-1};(\mu_l)_{l=1,\ldots,\bar{q}}\right]$ as

$$C_N^{\boldsymbol{\vartheta}}\equiv\left\{\boldsymbol{\vartheta}_0\colon\left(\hat{\boldsymbol{\vartheta}}_N-\boldsymbol{\vartheta}_0\right)'\left(\widehat{Var}\left(\hat{\boldsymbol{\vartheta}}_N\right)\right)^{-1}\left(\hat{\boldsymbol{\vartheta}}_N-\boldsymbol{\vartheta}_0\right)\leq c_{1-\alpha}\right\}.$$

It follows from the results in Rao (1973) (Section 7b) that

$$\lim_{N\to\infty}\Pr\left(\boldsymbol{\vartheta}\in C_N^{\boldsymbol{\vartheta}}\right)=1-\alpha.$$

Given $C_N^{\boldsymbol{\vartheta}}$, we can construct $C_N^{H[\mathbf{P}^x]}$ as follows. For a given $\boldsymbol{\vartheta}_0\in C_N^{\boldsymbol{\vartheta}}$, let $H_{\boldsymbol{\vartheta}_0}[\mathbf{P}^x]$ denote the identification region for $\mathbf{P}^x$ obtained when $\hat{\boldsymbol{\vartheta}}_N$ is replaced by $\boldsymbol{\vartheta}_0$ in the estimation procedure described in the previous section. Let

$$C_N^{H[\mathbf{P}^x]}=\bigcup_{\boldsymbol{\vartheta}_0\in C_N^{\boldsymbol{\vartheta}}}H_{\boldsymbol{\vartheta}_0}[\mathbf{P}^x].$$

Then

$$\boldsymbol{\vartheta}\in C_N^{\boldsymbol{\vartheta}}\Longrightarrow H[\mathbf{P}^x]\subseteq C_N^{H[\mathbf{P}^x]},$$

and therefore

$$\lim_{N\to\infty}\Pr\left(H[\mathbf{P}^x]\subseteq C_N^{H[\mathbf{P}^x]}\right)\geq 1-\alpha.$$

The confidence sets presented in Section 4 are obtained using this procedure. Using similar procedures one can construct confidence regions for $H\{\tau[\mathbf{P}^x]\}$ and $H\{\Theta[\mathbf{P}^x]\}$, where again $\tau(\cdot)$ and $\Theta(\cdot)$ denote functionals of $P(x)$.

3 Analysis of the Identifying Power of Specific Restrictions on $\Pi^\star$

This Section analyzes in detail examples of restrictions on the matrix $\Pi^\star$ (which satisfy Assumptions C0-C1) coming from validation studies and theories developed in the social sciences. I suggest settings in which such assumptions may be credible, show their implications for the structure of $H[\Pi^\star]$, and present results on the inferences that they allow one to draw on $\mathbf{P}^x$ and $\tau[\mathbf{P}^x]$. While the identification regions can be calculated and consistently estimated using the nonlinear programming method described in the previous section, it is often not possible to express them in closed form, unless $J=2$. Yet it is possible to derive closed form results for $H[\Pr(x=j)]$, $j \in X$, when the "base-case" assumptions are maintained. I will use these results as benchmark to evaluate the identifying power of additional assumptions. Notice however that $H[\Pr(x=j)]$, $j \in X$, is just the projection of $H[\mathbf{P}^x]$ on its $j-$th component. Hence, when $J>2$, a comparison based simply on $H[\Pr(x=j)]$, $j \in X$, understates the identifying power of the additional assumptions. When $J=2$, $H[\mathbf{P}^x]$ is entirely described by $H[\Pr(x=1)]$ and closed form bounds can be derived under different sets of assumptions, hence allowing for a full comparison.

3.1 Upper Bound on the Probability of Data Errors

Suppose that the researcher has a known lower bound on the probability that the realizations of w and x coincide, i.e., $\Pr(w=x) \geq 1-\lambda$, or, strengthening this assumption, that the researcher has a known lower bound on the probability of correct report for each value that x can take, i.e., $\Pr(w=j|\, x=j) \geq 1-\lambda$, $\forall j \in X$. Formally, consider the following:

Assumption 1 $\Pr(w=x) \geq 1-\lambda > 0$,

or, as a stronger version of Assumption 1, that

Assumption 2 $\Pr(w=j|\, x=j) \geq 1-\lambda > 0$, $\forall\, j \in X$.

Assumptions 1 and 2 are quite often satisfied in practice, mainly due to the availability of results of validation studies, and are therefore of particular interest. Additionally, as shown in Appendix A, Assumptions 1 and 2 exhaust the implications for the structure of $\Pi^\star$ of the assumptions typically maintained by researchers adopting mixture models. As already discussed, often the researcher has more or alternative information about the misreporting pattern than what is assumed in mixture models. Hence, the results obtained under these "base-case" assumptions are particularly suited to evaluate the identifying power of the available additional information. In the next section I will show that informative identification regions might be obtained even if one dispenses of Assumptions 1 and 2, when other information is available.

When the researcher has prior information suggesting that either Assumption 1, or the stronger Assumption 2, hold, she can specify the set $H^{E}[\Pi^{\star}]$, respectively, as follows:

$$H^{E,1}[\Pi^{\star}] = \left\{\Pi : \textstyle\sum_{h=1}^{J} \pi_{hh} p_h^x \geq 1-\lambda\right\},$$

$$H^{E,2}[\Pi^{\star}] = \{\Pi : \pi_{jj} \geq 1-\lambda,\ \forall j \in X\}.$$

where $H^{E,1}[\Pi^{\star}]$ denotes the set $H^{E}[\Pi^{\star}]$ when Assumption 1 is maintained, and $H^{E,2}[\Pi^{\star}]$ denotes the set $H^{E}[\Pi^{\star}]$ when Assumption 2 is maintained. Notice that $H^{E,2}[\Pi^{\star}] \subset H^{E,1}[\Pi^{\star}]$. Proposition 3 gives closed form bounds on $\Pr(x=j)$, $j \in X$, for the case in which either Assumption 1 or Assumption 2 holds.

Proposition 3 *a) Suppose that Assumption 1 holds, and that no other information is available. Then from system (1.1) we can learn that*

$$H[\Pr(x=j)] = \left[\max\left(P_j^w - \lambda, 0\right), \min\left(1, P_j^w + \lambda\right)\right],\ j \in X. \tag{3.1}$$

b) Suppose that Assumption 2 holds, and that no other information is available. Then from system (1.1) we can learn that

$$H[\Pr(x=j)] = \left[\max\left(\frac{P_j^w - \lambda}{1-\lambda}, 0\right), \min\left(1, \frac{P_j^w}{1-\lambda}\right)\right],\ j \in X. \tag{3.2}$$

□

The proof of Proposition 3 proceeds in two steps. First, it is shown that from the $j-$th equation of system (1.1) we can learn, depending on the maintained assumption, that $\Pr(x=j)$ lies in one of the intervals in (3.1)-(3.2). Then it is shown that there exists a $\Pi \in H[\Pi^{\star}]$ for which the extreme values of these intervals solve system (1.1), and that there exists no $\Pi \in H[\Pi^{\star}]$ for which a smaller lower bound or a bigger upper bound can be feasible. This implies that the bounds are sharp. The proof shows that when only Assumption 1 or Assumption 2 is maintained, only the $j-$th equation in system (1.1) implies restrictions on $\Pr(x=j)$, $j \in X$. In the next Section I will show that when more structure is imposed on the matrix Π, several of the equations in system (1.1) imply restrictions on $\Pr(x=j)$, $j \in X$, and additional progress can be made.

The same identification regions as those in Proposition 3 were obtained by Horowitz and Manski (1995). They used a mixture model to study the problem of inference with corrupted and contaminated data, and assumed that a known lower bound is available on the probability that a realization of w is drawn from the distribution of x. Molinari (2003) shows that under Assumptions 1 and 2, the identification regions for parameters that respect stochastic dominance obtained using the direct misclassification approach are also equivalent to those obtained by Horowitz and Manski (1995). Those results, along with Proposition 3 and Proposition 9 in Appendix A, show that when the error-ridden data take values in a finite set, and all the prior information is that Assumption 1 or Assumption 2 holds, the direct misclassification approach is equivalent to Horowitz and Manski's (1995) approach for drawing inference on $\Pr(x = j)$, $j \in X$, and on features of the distribution of x that respect stochastic dominance.

3.2 Constant Probability of Correct Report

Consider the case that, conditional on the value of x, there is constant probability that x is correctly reported, for at least a subset of the values that x can take. Formally:

Assumption 3 $\Pr(w = j|\, x = j) = \pi^\star \geq 1 - \lambda \geq 0 \ \forall j \in \tilde{X} \subseteq X$,

where $\pi^\star$ is known only to lie in $[1-\lambda, 1]$, and λ is strictly less than 1 if a nontrivial upper bound on the probability of a data error is available.

There are various situations in which this assumption may be credible. For example, Poterba and Summers (1995) use CPS data (with Reinterview Survey) and provide evidence (for the reinterviewed sub-sample) that the rate of correct report of employment status is similar for individuals who are employed or not in the labor force ($\Pr(w = j|\, x = j) \simeq 0.99$), but much lower for individuals who are unemployed ($\Pr(w = j|\, x = j) \simeq 0.86$). Kane, Rouse, and Staiger (1999) provide evidence (Table 5, p. 18) that self report of educational attainment is correct with similar probabilities for individuals with no college, some college but no AA degree, and AA degree ($\Pr(w = j|\, x = j) \simeq 0.92$), and is higher for individuals with at least a bachelor degree ($\Pr(w = j|\, x = j) \simeq 0.99$). Assumption 3 may hold with $\tilde{X} = X$ when the misclassification is generated by specific types of interviewer recording errors. For example, the interviewer may sometime mark one box at random in the questionnaire. Additionally, in the special case of dichotomous variables, some have argued that the misreporting of health disability is independent from true disability status (see Kreider and Pepper (2004) for a discussion of this issue), or that the misreporting of workers' union status is independent from true union status (see Bollinger (1996) for a discussion of this issue). When this is the case, Assumption 3 holds.

In general, Assumption 3 does not place any restriction on $\Pr(w = i|\, x = j)$, $i \neq j, i, j \in X$, other than that the misreporting probabilities need to satisfy

$$\textstyle\sum_{i \neq j} \Pr(w = i|\, x = j) = 1 - \pi^\star, \ \forall j \in \tilde{X}$$

When $J = 2$, this implies that the two off-diagonal elements of $\Pi^\star$ are equal; hence the only unknown element of $\Pi^\star$ is $\pi^\star$.

Suppose first that $\tilde{X} \subset X$, and without loss of generality let $\tilde{X} \equiv \{1, 2, \ldots, h\}$, $2 \leq h < J$. When this is the case, equation (1.1) can be rewritten as

$$\begin{bmatrix} \pi^\star & \pi^\star_{12} & \ldots & \pi^\star_{1J} \\ \pi^\star_{21} & \pi^\star & \ldots & \pi^\star_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \pi^\star_{J1} & \pi^\star_{J2} & \ldots & \pi^\star_{JJ} \end{bmatrix} \begin{bmatrix} \Pr(x = 1) \\ \Pr(x = 2) \\ \vdots \\ \Pr(x = J) \end{bmatrix} = \begin{bmatrix} \Pr(w = 1) \\ \Pr(w = 2) \\ \vdots \\ \Pr(w = J) \end{bmatrix} \tag{3.3}$$

where $\pi^\star \geq 1-\lambda$ and, assuming that λ constitutes a uniform upper bound for all the misclassification probabilities, $\pi^\star_{ll} \geq 1 - \lambda$, $\forall\ l \in \left(X - \tilde{X}\right)$. Then $H^E[\Pi^\star]$ will be defined as

$$H^{E,3}[\Pi^\star] = \left\{\Pi : \pi_{jj} = \pi \geq 1 - \lambda,\ \forall j \in \tilde{X};\ \pi_{ll} \geq 1 - \lambda,\ \forall\ l \in \left(X - \tilde{X}\right)\right\}.$$

Let $H^3[\Pi^\star] = H^P[\Pi^\star] \cap H^{E,3}[\Pi^\star]$, where $H^P[\Pi^\star]$ was defined in (2.4). Then one can immediately calculate $H[\mathbf{P}^x]$ and $H\{\tau[\mathbf{P}^x]\}$ using the nonlinear programming method described in Section 2, with $H^E[\Pi^\star] = H^{E,3}[\Pi^\star]$.

It is natural to ask whether Assumption 3 does have identifying power. To answer this question, in this section I consider the case that the researcher has a nontrivial upper bound on the probability of data errors, i.e. that $\lambda < 1$, and compare the bounds on $\Pr(x = j)$, $j \in X$, derived in Proposition 3, equation (3.2), with the extreme points obtained using the nonlinear programming method, with $H^E[\Pi^\star] = H^{E,3}[\Pi^\star]$. In Section 3.4 I consider the case in which x and w are binary ($J = 2$), and show that Assumption 3 can have identifying power even when $\lambda = 1$.

Proposition 4 shows that if $P^w_i > 0$, for some $i \in \tilde{X} \backslash \{j\}$, the base case lower bound on $\Pr(x = j)$, $j \in \tilde{X}$, if informative, is never feasible when Assumption 3 (with $\tilde{X} \subset X$) is maintained; hence the lower bound on $\Pr(x = j)$, $j \in \tilde{X}$ under Assumption 3 is strictly greater than that in (3.2). For the case in which the base case upper bound on $\Pr(x = j)$, $j \in \tilde{X}$ is informative, Proposition 5 derives conditions under which such upper bound is not feasible when Assumption 3 (with $\tilde{X} \subset X$) is maintained, and shows that when those conditions are satisfied, this upper bound is strictly smaller than that in (3.2). When the base case lower and upper bounds (respectively) are not informative, also the bounds on $\Pr(x = j)$, for a certain $j \in X$, are not informative.

Proposition 4 *(a) Suppose that Assumption 3 holds, with $\tilde{X} \subset X$, and that $P^w_j > \lambda$. Then the lower bound on* $\Pr(x = j)$, $j \in \tilde{X}$, *is strictly greater than the base case lower bound in (3.2). The base case lower bound in (3.2) is the sharp lower bound for* $\Pr(x = k)$, $k \in \left(X - \tilde{X}\right)$.
(b) Suppose that Assumption 3 holds, with $\tilde{X} \subset X$, and that $P^w_j \leq \lambda$. Then the sharp lower bound on $\Pr(x = j)$, $j \in X$, *coincides with the base case lower bound in (3.2), and is equal to* 0. □

Proposition 5 *(a) Suppose that Assumption 3 holds, with* $\tilde{X} \subset X$*, and that* $0 < P_j^w < 1-\lambda$*. If* $\lambda \leq \frac{1}{2}$*, the upper bound on* $\Pr(x=j)$ *,* $j \in \tilde{X}$*, is strictly smaller than the base case upper bound in (3.2) if and only if*

$$\exists\ k \in \tilde{X} \backslash \{j\} : P_j^w + P_k^w > (1-\lambda) + P_j^w \frac{\lambda}{1-\lambda}. \tag{3.4}$$

If $\lambda > \frac{1}{2}$*, the upper bound on* $\Pr(x=j)$ *,* $j \in \tilde{X}$*, is strictly smaller than the base case upper bound in (3.2) if*

$$\exists\ k \in \tilde{X} \backslash \{j\} : P_k^w > \lambda. \tag{3.5}$$

The base case upper bound in (3.2) is the sharp upper bound for $\Pr(x=k)$ *,* $k \in \left(X - \tilde{X}\right)$*.*

(b) Suppose that Assumption 3 holds, with $\tilde{X} \subset X$*, and that* $P_j^w \geq 1-\lambda$*. Then the sharp upper bound on* $\Pr(x=j)$ *,* $j \in X$*, coincides with the base case upper bound in (3.2), and is equal to* 1*.*
□

The proofs of Propositions 4-5, parts (a), are based on showing that there is no $\Pi \in H^3[\Pi^\star]$ for which the lower bound in (3.2) for $\Pr(x=j)$, $j \in \tilde{X}$, solves system (3.3), and that when condition (3.4) or condition (3.5) is satisfied, there is no $\Pi \in H^3[\Pi^\star]$ for which the upper bound in (3.2) for $\Pr(x=j)$, $j \in \tilde{X}$, solve system (3.3). When the inference is on $\Pr(x=k)$, $k \in \left(X - \tilde{X}\right)$, we can find a $\Pi \in H^3[\Pi^\star]$ that allows for the base case bounds in (3.2) to solve system (3.3). The proofs of Propositions 4-5, parts (b), are based on showing that when the bounds on $\Pr(x=j)$, $j \in X$, in (3.2) are not informative, one can find values of $\Pi \in H^3[\Pi^\star]$ for which $p_j^x = 0$ and $p_j^x = 1$ solve system (3.3).

The results in Propositions 4-5 can be explained as follows: only a subset $\tilde{X}$ of the equations in system (1.1) are related between each other. Therefore, when drawing inference on $\Pr(x=j)$, $j \in X$, an improvement on the base case bound in (3.2) can be achieved only for $j \in \tilde{X}$. Consider now the case in which $\tilde{X} = X$. In this case the results of Propositions 4-5 apply directly, with X replacing $\tilde{X}$. Of course, the identifying power of Assumption 3 is the highest in this case. In particular, inspection of Proposition 4 suggests that the lower bound for $\Pr(x=j)$, $j \in X$, if informative, improves for all j when Assumption 3 is maintained with $\tilde{X} = X$.

A final consideration is relevant. Often the researcher might have prior information suggesting that Assumption 3 holds, but not exactly. That is, she might have prior information that the probability of correct report is only approximately constant: $\Pr(w=j|\, x=j) \approx \pi^\star, \forall\ j \in \tilde{X} \subseteq X$. Then it is natural to ask how much the probabilities of correct report can differ between each other, for the results of Propositions 4-5 to still hold. For ease of exposition, consider the identification of $\Pr(x=1)$, and let $\pi_{11} = \pi$.[7] Molinari (2003) shows that as long as $|\pi_{jj} - \pi_{11}| < \lambda, \forall\ j \in \tilde{X} \backslash \{1\}$,

[7] When drawing inference on $P(x=j)$, $j \in \tilde{X}$, we can always define $\pi_{jj} = \pi$, and look at π_{kk}, $k \in \tilde{X} \backslash \{j\}$, as deviations from π.

and $\tilde{X} \subset X$, or $\tilde{X} = X$, the results of Proposition 4 continue to hold. A similar condition is derived for the results of Proposition 5.

Example 6 in Section 3.4 illustrates the identifying power of Assumption 3, both for the case in which $\tilde{X} \subset X$ and $\tilde{X} = X$, by comparing the identification regions $H\left[\Pr\left(x = j\right)\right], j \in X$, $H\left[\mathbf{P}^x\right]$ and $H\left[E\left(x\right)\right]$ obtained using the nonlinear programming method with $H^E\left[\Pi^\star\right] = H^{E,3}\left[\Pi^\star\right]$ with those obtained when only Assumption 2 is maintained.

3.3 Monotonicity in Correct Reporting

Social psychology suggests that when survey respondents are asked questions relative to socially and personally sensitive topics, they tend to underreport socially undesirable behaviors and attitudes, and overreport socially desirable ones. This suggestion is often supported by validation studies. In the context of questions of the type described above, these studies often document that $\Pr\left(w = j|\, x = j\right) \geq \Pr\left(w = j+1|\, x = j+1\right), \forall j \in \tilde{X} \subset X$. This is the case for example when survey respondents are asked about their participation in welfare programs, and $j = 1$ indicates non participation, while $j = 2$ indicates participation, or when they are asked about their employment status, and $j = 1, 2$ indicates, respectively, employed or not in the labor force, while $j = 3$ indicates unemployed.

Suppose that the set $X \equiv \{1, 2, \ldots, J\}$ can be ordered according to the "social desirability" of the values that x can take, with $x = 1$ being the most desirable, and $x = J$ the least desirable. Suppose further that the researcher believes that there is monotonicity in correct reporting. Then she can maintain the following:

Assumption 4 $\Pr\left(w = j|\, x = j\right) \geq \Pr\left(w = j+1|\, x = j+1\right), \forall\, j \in X \backslash \{J\}$, $\Pr\left(w = J|\, x = J\right) \geq 1 - \lambda \geq 0$,

where λ is strictly less than 1 if a nontrivial upper bound on the probability of a data error is available. When this assumption holds, $H^E\left[\Pi^\star\right]$ will be defined as

$$H^{E,4}\left[\Pi^\star\right] = \left\{\Pi : \pi_{jj} \geq \pi_{(j+1)(j+1)},\ \forall\, j \in X \backslash \{J\}\,,\ \pi_{JJ} \geq 1 - \lambda\right\}.$$

Let $H^4\left[\Pi^\star\right] = H^P\left[\Pi^\star\right] \cap H^{E,4}\left[\Pi^\star\right]$, where $H^P\left[\Pi^\star\right]$ was defined in (2.4). Then we can calculate $H\left[\mathbf{P}^x\right]$ and $H\left\{\tau\left[\mathbf{P}^x\right]\right\}$ using the nonlinear programming method described in Section 2, with $H^E\left[\Pi^\star\right] = H^{E,4}\left[\Pi^\star\right]$.

We are now left to verify that Assumption 4 does have identifying power. To accomplish this, we again consider the case that $\lambda < 1$, and compare the results that we can obtain using the nonlinear programming method when Assumption 4 is maintained, with those of Proposition 3. In Section

3.4 I consider the case in which x and w are binary ($J = 2$), and show that Assumption 4 can have identifying power even when $\lambda = 1$.

Suppose that Assumption 4 holds. Proposition 6 shows that the base case lower bound in (3.2), when informative, is feasible for $\Pr(x = 1)$. However, for $j \in X \backslash \{1\}$ if $P_l^w > 0$ for some $l \in \{1, \ldots, j-1\}$, the base case lower bound in (3.2), when informative, is not feasible for $\Pr(x = j)$, and hence the lower bound under Assumption 4 is strictly greater than that in (3.2). Regarding the base case upper bound in (3.2), the same results as those in Proposition 5 hold, with $\tilde{X} = \{j, j+1, \ldots, J\}$. The proof of this Proposition derives almost directly from the proofs of Propositions 4-5.

Proposition 6 *Suppose that Assumption 4 holds.*
a) Let $P_j^w > \lambda$. Then if $j = 1$, the base case lower bound in (3.2) is the sharp lower bound for $\Pr(x = 1)$. *The lower bound for* $\Pr(x = j)$, $j \in X \backslash \{1\}$, *is strictly greater than the base case lower bound in (3.2). The result of Proposition 4, part (b), is unchanged.*
b) Let $0 < P_j^w < (1 - \lambda)$. Then the same results as in Proposition 5 hold, with $\tilde{X} = \{j, j+1, \ldots, J\}$. The result of Proposition 5, part (b), is unchanged. □

Example 6 in Section 3.4 illustrates the identifying power of Assumption 4, by comparing the identification regions obtained using the nonlinear programming method with $H^E[\Pi^\star] = H^{E,4}[\Pi^\star]$ with those obtained when only Assumption 2 is maintained.

3.4 Dichotomous Variables and Numerical Examples

When x and w are dichotomous variables, the identifying power of Assumption 3 and Assumption 4 can be more easily appreciated, since the bounds on $H[\mathbf{P}^x]$ can be derived explicitly. This section shows how. It then provides numerical examples of the identification regions obtained under Assumptions 2, 3 and 4, both for the case of $J = 2$ and $J = 3$.

Let $X \equiv \{1, 2\}$.[8] The problem of misclassification of a dichotomous variable has received much attention in the econometric, statistical, and epidemiological literature. It is in the context of misclassified dichotomous variables that most of the precedents to the use of restrictions on the misclassification probabilities take place.

To start, suppose that Assumption 3 hold. In the related literature it has often been assumed that $\Pr(w = 1|\, x = 2) = \Pr(w = 2|\, x = 1)$, and additionally that these misclassification probabilities are less than $\frac{1}{2}$ (see, e.g., Klepper (1988) and Card (1996)). Notice that with dichotomous

[8] In the literature on dichotomous variables the two values that x can take are usually denoted $\{0, 1\}$. Here I use $\{1, 2\}$ to maintain the same notation as in the previous sections, where I denoted $X \equiv \{1, 2, \ldots, J\}$, $2 \leq J < \infty$.

variables Assumption 3 implies that equation (1.1) can be rewritten as

$$\begin{bmatrix} \Pr(w=1) \\ \Pr(w=2) \end{bmatrix} = \begin{bmatrix} \pi^\star & 1-\pi^\star \\ 1-\pi^\star & \pi^\star \end{bmatrix} \begin{bmatrix} \Pr(x=1) \\ \Pr(x=2) \end{bmatrix}.$$

Hence, the identification region $H[\mathbf{P}^x]$ can be inferred from the identification region

$$H[\Pr(x=1)] = \left\{p_1^x : P_1^w = \pi \cdot p_1^x + (1-\pi)\cdot(1-p_1^x),\ \pi \in H^3[\Pi^\star]\right\}.$$

where $H^3[\Pi^\star]$ was defined in Example 2. Notice that if $\pi = \frac{1}{2}$, $P_1^w = \frac{1}{2}$; in this case, $P(w|\, x) = P(w)$, i.e. x and w are statistically independent, and obviously knowledge of $P(w)$ does not provide any information on $P(x)$. If $P_1^w \neq \frac{1}{2}$, we know that $\pi \neq \frac{1}{2}$. The following Proposition characterizes explicitly $H[\Pr(x=1)]$.

Proposition 7 *Let Assumption 3 hold, with $\tilde{X} = X \equiv \{1,2\}$.*
a) If $\lambda < \frac{1}{2}$, then

$$\begin{cases} H[\Pr(x=1)] = \left[P_1^w, \min\left(\frac{P_1^w-\lambda}{1-2\lambda}, 1\right)\right] & \textit{if } P_1^w \geq 0.5, \\ H[\Pr(x=1)] = \left[\max\left(\frac{P_1^w-\lambda}{1-2\lambda}, 0\right), P_1^w\right] & \textit{otherwise.} \end{cases}$$

b) If $\lambda \geq \frac{1}{2}$, then

$$\begin{cases} H[\Pr(x=1)] = [P_1^w, 1] & \textit{if } P_1^w > \lambda, \\ H[\Pr(x=1)] = \left[0, \frac{P_1^w-\lambda}{1-2\lambda}\right] \cup [P_1^w, 1] & \textit{if } \lambda \geq P_1^w \geq \frac{1}{2}, \\ H[\Pr(x=1)] = [0, P_1^w] \cup \left[\frac{P_1^w-\lambda}{1-2\lambda}, 1\right] & \textit{if } \frac{1}{2} > P_1^w \geq 1-\lambda, \\ H[\Pr(x=1)] = [0, P_1^w] & \textit{if } 1-\lambda > P_1^w. \end{cases}$$

These identification regions are a subset of those in (3.2). □

The fact that if $\lambda \geq \frac{1}{2}$, $H[\Pr(x=1)]$ can be given by two disjoint intervals is a direct consequence of the possible disconnectedness of $H[\Pi^\star]$ arising when one assumes constant probability of correct report, and described in Section 2 and in Example 2.

Suppose now that Assumption 4 hold. Also in this case the identification region $H[\mathbf{P}^x]$ can be inferred from the identification region

$$H[\Pr(x=1)] = \left\{p_1^x : P_1^w = \pi_{11} \cdot p_1^x + (1-\pi_{22})\cdot(1-p_1^x),\ (\pi_{11},\pi_{22}) \in H^4[\Pi^\star]\right\}, \tag{3.6}$$

where $H^4[\Pi^\star]$ was defined in Example 3. Notice that again if $\pi_{11} = \pi_{22} = \frac{1}{2}$, $P_1^w = \frac{1}{2}$; in this case, $P(w|\, x) = P(w)$, i.e. x and w are statistically independent, and obviously knowledge of $P(w)$ does not provide any information on $P(x)$. If $P_1^w \neq \frac{1}{2}$, we know that π_{11} and π_{22} cannot be jointly equal to $\frac{1}{2}$. The following Proposition characterizes explicitly $H[\Pr(x=1)]$.

Proposition 8 *Let Assumption 4 hold.*
a) If $\lambda < \frac{1}{2}$, then

$$\begin{cases} H\left[\Pr(x=1)\right] = \left[\max\left(\frac{P_1^w-\lambda}{1-\lambda},0\right), \min\left(\frac{P_1^w-\lambda}{1-2\lambda},1\right)\right] & \text{if } P_1^w \geq 0.5, \\ H\left[\Pr(x=1)\right] = \left[\max\left(\frac{P_1^w-\lambda}{1-\lambda},0\right), P_1^w\right] & \text{otherwise.} \end{cases} \tag{3.7}$$

b) If $\lambda \geq \frac{1}{2}$, then

$$\begin{cases} H\left[\Pr(x=1)\right] = \left[\frac{P_1^w-\lambda}{1-\lambda},1\right] & \text{if } P_1^w > \lambda \\ H\left[\Pr(x=1)\right] = [0,1] & \text{if } \lambda \geq P_1^w \geq \frac{1}{2} \\ H\left[\Pr(x=1)\right] = [0,P_1^w] \cup \left[\frac{P_1^w-\lambda}{1-2\lambda},1\right] & \text{if } \frac{1}{2} > P_1^w \geq 1-\lambda \\ H\left[\Pr(x=1)\right] = [0,P_1^w] & \text{if } 1-\lambda > P_1^w \end{cases} . \tag{3.8}$$

These identification regions are a subset of those in (3.2). □

Again, the fact that if $\lambda \geq \frac{1}{2}$ and $P_1^w < \frac{1}{2}$, $H\left[\Pr(x=1)\right]$ can be given by two disjoint intervals is a direct consequence of the possible disconnectedness of $H\left[\Pi^\star\right]$ arising when one assumes monotonicity in correct reporting, and described in Section 2 and in Example 3.

The following numerical example illustrates the identifying power of Assumption 3 and Assumption 4, with $X = \{1,2\}$, by comparing the bounds in Propositions 7 and 8 with those in (3.2), and showing how the bounds improve as λ gets closer to the true misclassification parameter.

Example 5 *Let $\Pr(x=1) = 0.3$, and $\pi^\star = 0.9$, so that $P_1^w = 0.34$. Table 1 gives lower and upper bounds on $\Pr(x=1)$, when Assumptions 2, 3 and 4 are maintained, as λ approaches $1-\pi^\star$. Notice that the identification region for $\Pr(x=1)$, when Assumptions 3 and 4 are maintained, is informative even when $\lambda = 1$.*

To conclude this section, I illustrate the identifying power of Assumption 3 (both for the case in which $\tilde{X} \subset X$ and $\tilde{X} = X$) and Assumption 4, when $J = 3$. I compare the identification regions $H\left[\Pr(x=j)\right]$, $j \in X$, $H\left[\mathbf{P}^x\right]$ and $H\left[E(x)\right]$ obtained using the nonlinear programming method with $H^E\left[\Pi^\star\right] = H^{E,3}\left[\Pi^\star\right]$ and with $H^E\left[\Pi^\star\right] = H^{E,4}\left[\Pi^\star\right]$ with those obtained when only Assumption 2 is maintained.

Example 6 *Let: $X = \{1,2,3\}$, $\lambda = 0.2$, $\pi^\star = 0.85$, $\left[\Pr(x=j), j \in X\right] = \left[0.3 \;\; 0.6 \;\; 0.1\right]^T$, and suppose that $\pi_{21}^\star = 0.11$, $\pi_{12}^\star = 0.13$, $\pi_{13}^\star = 0.04$, so that $\mathbf{P}^w = \left[0.34 \;\; 0.55 \;\; 0.11\right]^T$; with these values, $E(x) = 1.8$. Table 2 gives the identification regions for $\tau\left[\mathbf{P}^x\right] = \Pr(x=j)$, $j \in X$, and for $\tau\left[\mathbf{P}^x\right] = E(x)$, when Assumption 2 alone is maintained, when Assumptions 2 and 3 are* *jointly maintained with $\tilde{X} = X$ and with $\tilde{X} = \{1, 2\}$, and when Assumptions 2 and 4 are jointly maintained. The improvement in the upper bound on $\Pr(x = 1)$ comes from the second equation of system (1.1); indeed $P_1^w + P_2^w = 0.89 > 0.885 = (1 - \lambda) + \frac{\lambda}{1-\lambda} P_1^w$. Figure 2 plots the identification regions $H[\mathbf{P}^x]$ obtained under the different assumptions.*

4 Estimation and Inference for the Distribution of Pension Plan Types in the U. S.

To illustrate estimation of the bounds and construction of the confidence sets, I consider data on the distribution of pension plan characteristics in the American population age $51 - 61$. The data are based on household interviews obtained in the Health and Retirement Study (HRS), a longitudinal, nationally representative study of older Americans, which in its base year of 1992 surveyed $12,652$ individuals from $7,607$ households, with at least one household member born between 1931 and 1941. The survey has been updated every two years since 1992, and in 1998 a new cohort of $2,529$ individuals born between 1942 and 1947 (so called "War Babies") was added to the HRS sample. I use data from the first HRS wave and from the War Babies wave, focusing on the information collected on pension plan characteristics for people age $51 - 61$ and employed at the time of the survey. This provides two nationally representative cross-sections of the population of interest. The question to be addressed is:

> How did the distribution of pension plan types in the population of currently employed Americans, age $51 - 61$, change between 1992 and 1998?

Three pension plan types are possible: defined benefit (DB), defined contribution (DC), and plans incorporating features of both (Both). Defined benefit and defined contribution plans differ greatly in their characteristics. As described by Gustman, Mitchell, Samwick, and Steinmeier (2000), in a defined benefit pension the benefit formula is specified by the plan sponsor, usually as a function of the worker's highest salary, years of service, and retirement age. After an initial period, the worker gains a right to an eventual pension benefit at the plan's retirement age. Typically such plans reduce the benefit amount for retirement prior to the so-called normal retirement age. DB plans are usually financed by employer (pre-tax) contributions. On the other hand, DC plans do not specify the retirement benefit, but they set how much will be contributed into the account each year the worker remains with the plan. Then the benefit payout is determined at retirement, as a function of how much it accumulated in the worker's account. The plan type can affect several pension-related variables, including pension wealth and pension accrual, that is, the change in pension wealth when a worker delays retirement by one year. For example, there are DB plans in which an additional year of service is rewarded by greater retirement benefits up to the firm's early retirement age. Then the benefit accrual profile may flatten out, and even become negative, if retirement is delayed further. By contrast, DC plans tend to be actuarially neutral with regard to the retirement age, rewarding delayed retirement more monotonically.

It is then of interest to learn how the distribution of pension plan types has changed over time, as a preliminary step before studying the relation between pension incentives and retirement and saving behavior. The HRS data can provide valuable information in this direction. However, there is evidence that workers are particularly misinformed about their pension plans' characteristics, and it is therefore not obvious how to make use of their reported pension plans' description to draw the inference of interest. Gustman and Steinmeier (2001) linked data from the first HRS wave with restricted data from Social Security Administration and employer provided pension plan description, and documented that individuals with matched data (approximately 51% of the entire HRS sample, and 67% of currently employed respondents) approaching retirement age are remarkably misinformed with regard to their pension plans' characteristics. Their results are reported in Table 3, and suggest that overall, approximately 49% of the currently employed individuals with matched data correctly identify their pension plan type, the remaining 51% providing a wrong report.

For the individuals in the first HRS wave without a matched pension (33% of the sample) it is difficult to determine the true plan type: on one side, Gustman and Steinmeier (2001) document that the sub-sample without a matched pension is different from the sub-sample with a matched pension; on the other side, the evidence for the sub-sample with matched pension casts doubts on the reliability of the self reports. Moreover, linked data are not available for individuals in subsequent waves, or for individuals in the War Babies wave.[9] Yet, the results of Gustman and Steinmeier's (2001) analysis provide information on the misreporting pattern, and such information can be exploited through the direct misclassification approach to draw inference on how the distribution of pension plan types for the population as a whole has changed between 1992 and 1998, using data from the first HRS wave and from the War Babies wave.

In all that follows I will assume that the HRS respondents correctly report whether they are covered by a pension,[10] and I will take firm reported plan types to be the "true" plan types. Let $x = 1$ if the individual has a DB plan, $x = 2$ if the individual has a DC plan, and $x = 3$ if the individual has a plan combining features of both, so that $X \equiv \{1, 2, 3\}$. As before, $w \in X$ denotes the reported pension plan type. Let $\mathbf{P}^{w,t} \equiv [\Pr_t(w = j), j \in X]$ and $\mathbf{P}^{x,t} \equiv [\Pr_t(x = j), j \in X]$

[9] Additionally, employer provided pension plan descriptions are not publicly accessible by HRS users. In particular, such data are not available for the analysis carried out in this paper.

[10] This assumption is based on Gustman and Steinmeier's (2001) comparison between peoples' report on their pension coverage in both the 1992 and 1994 waves of the HRS. This comparison shows that 93% of the respondents who declared to be covered by a pension or to be not covered by a pension in 1992, give the same answer in 1994. Of the remaining 7%, approximately 80% are individuals who declared not to be covered by a pension in 1992, but to be covered in 1994.

denote, respectively, the vectors of fractions of reported pension plan types and true pension plan types at time $t = 1992, 1998$. For the respondents in the first HRS wave, let $s_l = 1$ denote the fact that individual $l \in L_{1992}$ has a matched pension plan description, $s_l = 0$ otherwise, and denote by $\Pi^{\star 1}_{1992}$ the matrix of misclassification probabilities that maps the true pension plan types into the reported types for individuals with matched pension plan descriptions. Let $\Pi^{\star 0}_{1992}$ denote the matrix of misclassification probabilities for the respondents in the first HRS wave without a matched plan description, and let $\Pi^{\star}_{1998}$ denote the matrix of misclassification probabilities for the entire sample of respondents in the War Babies wave. Table 3 reveals, up to statistical considerations, $\Pi^{\star 1}_{1992}$. From the HRS data and from Gustman and Steinmeier's (2001) results we can learn $\mathbf{P}^{w,1992}$, $\mathbf{P}^{w,1998}$, and $[\Pr_{1992}(x = j|\, s = 1), j \in X]$. These values are reported in Table 4, along with 95% bootstrap confidence intervals.

One might expect the misclassification pattern reported by Gustman and Steinmeier (2001) to hold also for the subset of respondents without matched pension plan descriptions. On the other hand, one might expect that the misclassification structure mapping true pension plan types into reported types changes over time, so that $\Pi^{\star 1}_{1992}$ can help in constructing $H[\Pi^{\star}_{1998}]$, but not reduce this set to a singleton. However, one might as well be tempted to entertain assumptions strong enough to achieve point identification of the quantity of interest. To test the credibility of these conjectures, I will examine the following assumptions:

Assumption E1: No Selection. $\Pi^{\star 0}_{1992} = \Pi^{\star 1}_{1992}$.

Assumption E2: No Selection and No Variation Over Time. $\Pi^{\star}_{1998} = \Pi^{\star 1}_{1992}$.

The first assumption states that the misreporting pattern is the same across respondents in the first HRS wave with matched pension plan description and without matched pension plan description. The second assumption states that the misreporting pattern for the respondents in the War Babies wave is the same as that for the respondents with matched data in the first HRS wave. When these assumptions are maintained, $\Pi^{\star}_{1992}$ and $\Pi^{\star}_{1998}$ are identified, and, since $\Pi^{\star 1}_{1992}$ is nonsingular, one can use the equation $\mathbf{p}^x = \Pi^{-1} \cdot \mathbf{P}^w$ to attempt to learn $[\Pr_t(x = j), j \in X]$, $t = 1992, 1998$. Table 5 reports the results of such procedure, along with 95% bootstrap confidence intervals. As we can see from the table, the data reject the assumption that $\Pi^{\star}_{1998} = \Pi^{\star 1}_{1992}$: the vector obtained from solving $\left(\Pi^{\star 1}_{1992}\right)^{-1} \cdot \mathbf{P}^{w,1998}$ does not generate a valid probability measure. In particular, the first element of the implied vector is negative, and its 95% confidence interval does not cover the zero, and the last element is greater than one. Hence, point identification of $\mathbf{P}^{x,1998}$ through Assumption E2 is not possible. On the other hand, the data do not reject the assumption that $\Pi^{\star 0}_{1992} = \Pi^{\star 1}_{1992}$, despite the possible selection problem. In all that follows I will maintain Assumption E1 and focus the attention on the problem of inferring $H\left[\mathbf{P}^{x,1998}\right]$. Of course, Assumption E1 can be relaxed, and $H\left[\mathbf{P}^{x,1992}\right]$ can be estimated under weaker assumptions using the direct misclassification approach.

The main assumption that I will maintain throughout the entire analysis, and that I use to exploit part of the information in $\Pi^{\star 1}_{1992}$ to learn $H\left[\mathbf{P}^{x,1998}\right]$, is the following:

Assumption E3: No Reduction in Awareness. $\pi_{jj,1998} \geq \pi_{jj,1992}$, $\forall\ j \in X$.

This assumption amounts to say that the fraction of individuals correctly identifying their pension plan type does not decline over time. This in turn implies that lower bounds on the probability of correct report in 1992 provide lower bounds on the probability of correct report in 1998. Assumption E3 is motivated by the observation that in recent years the Social Security Administration and the Department of Labor have increasingly expanded their efforts to improve individuals' knowledge about pensions and about retirement saving in general (see Gustman and Steinmeier (2001) for a summary of recent interventions).

I now introduce two sets of assumptions, which I entertain along with Assumption E3 to construct the set $H\left[\Pi^{\star}_{1998}\right]$, and derive $H\left[\mathbf{P}^{x,1998}\right]$. Of course, different empirical researchers might hold disparate beliefs about which of the assumptions in Cases 1 and 2 hold, and moreover they might bring to bear different prior information. However, the results of the analysis are interesting both in that they show the functioning of the direct misclassification approach, as well as in that they shed some light on the question of interest. The goal of the analysis is to learn the change in the fraction of individuals in the US population approaching retirement age having a DB plan.

The identification regions that I obtain for $H\left[\mathbf{P}^{x,1998}\right]$ are plotted in Figure 3, along with their 95% Confidence Sets. The identification regions $H\left[\Pr_{1998}\left(x=j\right)\right], j \in X$, are reported in Table 6, again with their 95% confidence intervals.

<u>**Case 1:**</u>

$$H\left[\Pi^{\star}_{1998}\right] = H^{P}\left[\Pi^{\star}\right] \cap \left\{\Pi : \pi_{11} \approx \pi_{22} \geq 0.53,\ \pi_{22} \geq \pi_{33} \geq 0.34,\ \pi_{21} \leq \pi_{12},\ \pi_{31} \leq \pi_{13},\ \pi_{23} \leq \pi_{13}\right\}.$$

Case 1 maintains Assumption E3, and builds on Assumption E1. Jointly, these assumptions imply that the same pattern of correct report as observed for $\Pi^{\star}_{1992}$ holds also for the sample of respondents in the War Babies wave, hence providing lower bounds on the probabilities of correct report. Additionally, I also require constant probability of correct report for individuals who truly have DB and DC plans. This assumption is motivated by observing, in Table 3, that $\Pr\left(w=1|\, x=1, s=1\right) \approx \Pr\left(w=2|\, x=2, s=1\right)$. Finally, I make monotonicity assumptions on some of the misclassification probabilities. In particular, Table 3 suggests that individuals who truly have a plan incorporating features of both DB and DC classify their plan into the category of DB plans much more often than individuals with DB plans report plans incorporating features of both (0.45 vs. 0.27). Similarly, individuals who truly have a DC plan report a DB plan more often than individuals with a DB plan report a DC one (0.26 vs. 0.15). Also, individuals who truly have a plan incorporating features of both DB and DC report a DB plan more often than a DC one (0.45 vs. 0.18). This seems to reveal a tendency of respondents to remarkably misreport in the direction of DB plans; such tendency is incorporated in assuming $\pi_{21} \leq \pi_{12}$, $\pi_{31} \leq \pi_{13}$, $\pi_{23} \leq \pi_{13}$.

The first panel of Figure 3 shows the estimate of $H\left[\mathbf{P}^{x,1998}\right]$ obtained in Case 1. It is interesting to observe that the estimated set displays nonconvexities, a feature that the nonlinear programming estimator is capable to capture. The third panel of the figure displays the 95% confidence set of $H\left[\mathbf{P}^{x,1998}\right]$. For the construction of this confidence set, I estimated $\mathbf{P}^{w,1998}$ using sample means, and took as estimates of the lower bounds in $H^E\left[\Pi^\star\right]$ the values $\mu_{1,n}$, $\mu_{2,n}$ in the (2,2) and (3,3) entries of Table 3. While borrowed from Gustman and Steinmeier (2001), these estimates are based on a validation data (respondents to the 1992 wave with matched pension plan descriptions) independent from the 1998 data, and with $n = 2,907$. For the construction of the confidence ellipsoid for $\left[P_1^{w,1998}, P_2^{w,1998}, \mu_1, \mu_2\right]$ I used $\kappa = \frac{N}{n} = \frac{1{,}124}{2{,}907}$. The estimates of $\Pr_{1992}(x=1)$ and $H\left[\Pr_{1998}(x=1)\right]$ reported in Table 6 suggest that the fraction of individuals having a DB plan should have declined between 1992 and 1998. However, the confidence intervals of the two estimates do overlap; hence we cannot reject the hypothesis $\Pr_{1992}(x=1) - \Pr_{1998}(x=1) < 0$. This shows that under relatively mild restrictions we can obtain a strong conclusion regarding our question of interest, although more assumptions are needed to obtain statistical significance.

Case 2:

$$H\left[\Pi_{1998}^\star\right] = H^P\left[\Pi^\star\right] \cap \left\{ \Pi : \left(\begin{array}{l} \pi_{11} \approx \pi_{22} \geq \pi_{33} \geq 0.53, \\ \pi_{21} \leq \pi_{12}, \pi_{31} \leq \pi_{13}, \pi_{23} \leq \pi_{13}, \\ \pi_{21} \geq 0.10, \pi_{ij} \geq 0.15 \text{ for all other } i,j \in X, i \neq j. \end{array} \right) \right\}$$

Case 2 builds on Case 1, as it retains all the assumptions maintained there. However, it is crucially set apart from the previous case, in that it requires a lower bound on each probability of misclassification. This in turn implies that, given any true pension plan type, the probability of correct report has to be necessarily less than one. This assumption is motivated by the large amount of misreporting of pension plan types which appears in Table 3, and which is documented at large by Gustman and Steinmeier (2001). Additionally, π_{33} is required to have the same lower bound as π_{11} and π_{22}. This is motivated by the large amount of information campaigns on DC plans (in particular 401k) that has characterized the mid to late 1990s.

Under these assumptions, the estimate of $H\left[\mathbf{P}^{x,1998}\right]$ shrinks further. This allows one to conclude that the fraction of individuals having DB plans has decreased between 1992 and 1998; in particular, $\Pr_{1992}(x=1) - \Pr_{1998}(x=1) \geq 0.14$. This in turn implies that the fraction of individuals having either DC plans or plans incorporating features of both has increased sharply between 1992 and 1998. While the confidence intervals for the parameters of interest do not overlap, so that the assumption $\Pr_{1992}(x=1) - \Pr_{1998}(x=1) < 0$ can be rejected, we cannot reject the assumption $\Pr_{1992}(x=1) - \Pr_{1998}(x=1) = \beta$ for values of β in $[0.06, 0.5]$. The confidence set for Case 2 is constructed again by estimating $\mathbf{P}^{w,1998}$ using sample means, and taking as estimate of the lower bound for π_{jj}, $j = 1,2,3$, in $H^E[\Pi^\star]$ the value μ_n in the (2,2) entry of Table 3. However the lower bounds for the other parameters are treated as constant, so that the confidence ellipsoid is constructed exclusively for the vector $\left[P_1^{w,1998}, P_2^{w,1998}, \mu\right]$.

5 Extensions

The direct misclassification approach can be easily extended to drawing inference in presence of multiple misclassified variables, regression with misclassified outcome, regression with misclassified regressor, and jointly missing and misclassified outcomes. Below I list briefly the modifications of the approach that will allow inference in each of these cases.

1. Two or More Misclassified Variables.

In this case, the researcher will simply have to redefine variables. Suppose that interest centers on features of $P\left(x^1, x^2\right)$, $x^1 \in X^1 \equiv \{1,2,\ldots,J_1\}$, $x^2 \in X^2 \equiv \{1,2,\ldots,J_2\}$, $2 \leq J_1, J_2 < \infty$, and the researcher observes only $\left(w^1, w^2\right)$, a misclassified version of $\left(x^1, x^2\right)$. She can then construct random variables s and r, taking values in $S \equiv \{1,2,\ldots,J_1 \cdot J_2\}$, and such that $s = (l-1)\cdot J_2 + j$ if $x^1 = j$ and $x^2 = l$, and $r = (k-1)\cdot J_2 + i$ if $w^1 = i$ and $w^2 = k$. She can then write the analogue of equation (1.1) for r and s, and use the method proposed here to draw the inference of interest.

2. Regressions.

(a) If interest centers on features of $P\left(x|\, s = \mathrm{s}_0\right)$, where $s \in S$ is a perfectly observable discrete covariate with $\Pr\left(s = \mathrm{s}_0\right) > 0$, and the researcher has prior information on $\Pi^\star_{\mathrm{s}_0} \equiv \{\Pr\left(w = i|\, x = j, s = \mathrm{s}_0\right)\}_{i,j\in X}$, the proposed method can be applied directly, with the event $s = \mathrm{s}_0$ conditioning all the probabilities involved.

(b) Consider now the case that interest centers on features of $P\left(y|\, x\right)$, where y is a perfectly observed outcome variable. The problem of regression with misclassified covariates has been widely studied (e.g., Aigner (1973), Klepper (1988), Bollinger (1996), Card (1996), Kane, Rouse, and Staiger (1999), Hu (2003), Mahajan (2003)), and point identified or interval identified estimators have been proposed under specific sets of assumptions. The direct misclassification approach can be used to estimate the smallest point and the largest point in the identification region of (for example) a mean regression under any set of assumptions. Molinari (2003) shows how. Here I present ideas, for the special case in which the probability of correct report is greater than $\frac{1}{2}$ for each of the values that x can take (and any additional assumption might hold). In this case we already discussed that any $\Pi \in H[\Pi^\star]$ is of full rank, so that $\mathbf{p}^x = \Pi^{-1}\cdot\mathbf{P}^w$. This implies that $P\left(x|\, w\right)$ can be uniquely expressed as a function of Π. First, suppose that $H[\Pi^\star]$ is a singleton, so that $P\left(w|\, x\right)$ is identified, and therefore $P(x)$ and $P\left(x|\, w\right)$ are identified as well. $P\left(y|\, w, x\right)$ and $P\left(y|\, x\right)$ remain unknown, but knowledge of $P(y|\,w)$ and $P(x|\,w)$ imply restrictions on $[P(y|\,w=i,x=j)\,,\ i,j\in X]$. Hence, for any $i\in X$, we can draw inference on $E(y|\,w=i,x=j)$, $j\in X$, and then use this information, knowledge of $P(w|\,x)$, and the Law of Total Probability to draw inference on $E(y|\,x)$. In particular, from the entire population, consider the sub-population with $w=i$. Horowitz and Manski (1995) showed that the smallest feasible value of $E(y|\,w=i,x=j)$ occurs if, within this sub-population, the persons with $x=j$ have the smallest values of y. Similarly, they showed that the largest feasible value occurs if the persons with $x=j$ have the largest values of y.[11] The smallest value of $E(y|\,x=j)$ will then be given by the weighted sum of the smallest values of $E(y|\,w=i,x=j)$ obtained for each $i\in X$, with weights given by $\pi^{\star}_{ij}$. Similarly, the largest value of $E(y|\,x=j)$ will be given by the weighted sum of the largest values of $E(y|\,w=i,x=j)$ obtained for each $i\in X$, again with weights given by $\pi^{\star}_{ij}$. Consider now the (general) case in which $H[\Pi^{\star}]$ is not a singleton, so that $P(w|\,x)$, and therefore $P(x)$ and $P(x|\,w)$, are not identified. For given $\Pi\in H[\Pi^{\star}]$, $\mathbf{p}^{x}\in H[\mathbf{P}^{x}]$ and a feasible value of $[\Pr(x=j|\,w=i)\,,i,j\in X]$ are determined. Hence, for each $\Pi\in H[\Pi^{\star}]$, one can repeat the same argument as that above, and express the largest and the smallest points in the identification regions for $E(y|\,x=j)$ (derived for each $\Pi\in H[\Pi^{\star}]$) as functions of Π. Taking the infimum and the supremum, respectively, of these smallest and largest points for $\Pi\in H[\Pi^{\star}]$ gives the smallest and the largest point in $H[E(y|\,x=j)]\,,\ j\in X$.

This same argument has been proposed by Dominitz and Sherman (2003), who studied the problem of inferring the distribution of test scores for truly English proficient students ($x=1$), when only an imperfect indicator of English proficiency is available ($w=1$). They used a mixture model with verification, and assumed that students classified as English proficient ($w=1$) are more likely to be truly English proficient ($x=1$) than students classified as limited English proficient ($w=2$). In terms of misclassification probabilities, this assumption translates into $\pi_{11}\geq P^{w}_{1}$.

3. Jointly Missing and Misclassified Data.

The data available to the empirical researcher are often not only error ridden, but also incomplete. Consider the example of survey respondents being asked about their pension plan type: not only they can report DB, DC, or Both, but they can as well choose not to respond to the question. Let $w=J+1$ denote this outcome. Then system (1.1) can be rewritten as follows:

$$\begin{bmatrix} \Pr(w=1) \\ \vdots \\ \Pr(w=J) \\ \Pr(w=J+1) \end{bmatrix} = \begin{bmatrix} \Pr(w=1|\,x=1) & \ldots & \Pr(w=1|\,x=J) \\ \vdots & \ddots & \vdots \\ \Pr(w=J|\,x=1) & \ldots & \Pr(w=J|\,x=J) \\ \Pr(w=J+1|\,x=1) & \ldots & \Pr(w=J+1|\,x=J) \end{bmatrix} \begin{bmatrix} \Pr(x=1) \\ \vdots \\ \Pr(x=J) \end{bmatrix}.$$

[11] Denoting by $r_k(\cdot)$ the quantile function corresponding to $P(y|\,w=k)$, $k\in X$, these smallest and largest values of $E(y|\,w=i,x=j)$ correspond to the expectations of the observable distribution $P(y|\,w=i)$, respectively right truncated at $r_i(\Pr(x=j|\,w=i))$ and left truncated at $r_i(1-\Pr(x=j|\,w=i))$.

This simply implies that the set $H[\Pi^{\star}]$ is a set of rectangular matrices. The identification regions $H[\mathbf{P}^x]$ and $H\{\tau[\mathbf{P}^x]\}$ are still defined as in (2.2)-(2.3), and the nonlinear programming method can be used to consistently estimate them. Of course, there will be additional constraints, one coming from the $(J+1)-$th equation in the above system, and the others from possible assumptions on the relationship between misreporting and nonresponse. However the direct misclassification approach can still be used to draw the inferences of interest.

6 Conclusions

This paper has studied the problem of drawing inference when a discrete variable is subject to classification errors. This is a commonplace problem in surveys and elsewhere. The problem has long been conceptualized through *convolution* and *mixture models.* This paper introduced the *direct misclassification approach.* The approach is based on the observation that in the presence of classification errors, the relation between the distribution of the "true" but unobservable variable and its misclassified representation is given by a linear system of simultaneous equations, in which the coefficient matrix is the matrix of misclassification probabilities.

While this matrix is unknown, validation studies, economic theory, cognitive and social psychology, or knowledge of the circumstances under which the data have been collected can provide information on the misclassification pattern that has transformed the "true" but unobservable variable into the observable but possibly misclassified variable. The method introduced in this paper shows how to transform such prior information into sets of restrictions on the (unknown) matrix of misclassification probabilities, and exploit these restrictions to derive identification regions for any real functional of the distribution of interest, using the linear system of simultaneous equations directly. By contrast, mixture models do not allow the researcher to easily exploit this type of prior information to learn features of the distribution of interest. Convolution models, as usually implemented with the assumption of independence between measurement error and "true" variable, are not suited to analyze errors in discrete data. The direct misclassification approach does not rely on any specific set of assumptions, but it can incorporate into the analysis any prior information that the researcher might have on the misreporting pattern. In some cases the implied identification regions have a simple closed form solution, that allows for straightforward estimation using sample analogs. When this is not the case, the identification regions can be estimated using the nonlinear programming estimator introduced in this paper. Confidence sets that cover the true identification region with probability at least equal to a prespecified confidence level can be constructed using a simple procedure based on the inversion of a Wald statistic.

A *Misclassification Models* and *Mixture Models*

Due to the pervasiveness of the problem, inference in the presence of error-ridden data has been widely studied both in statistics and econometrics. Rather than focusing on equation (1.1) directly, each of these fields has conceptualized the problem through two main models: *mixture models* and *convolution models*. In what follows I show that, in the specific case of variables taking values in a finite set $X \equiv \{1, 2, \ldots, J\}$, $2 \leq J < \infty$, these models can be formally expressed as *misclassification models*.

When an analyst adopts a convolution model, she generally believes that the variable of interest is affected by "chronic errors", i.e. that the error distributions have no mass point at zero. She then views the available data as realizations of a random variable w which measures the unobservable x with errors in variables:

$$w \equiv x + v,$$

where v is a random variable which represents the imperfection in the measurement of x. In this case the relation between the observable distribution of w and the unobservable distribution of x is given by

$$\begin{aligned} P(w) &= P(x+v), \\ P(x) &= P(w-v). \end{aligned}$$

The analyst will assume that x and v are uncorrelated, or even independent, and that $E(v) = 0$.

When a variable with finite support is imperfectly classified, the assumption of independence between measurement error and true variable cannot hold. Moreover, validation studies suggests that a significant part of the observed data are error free. In terms of a convolution model, this means that the error distribution has a mass point at zero. Once we introduce the mixture model, it will become apparent that if this is the case, the convolution model can be treated as a mixture model, and the results derived for the mixture model apply thoroughly.

When an analyst adopts a mixture model, she implicitly or explicitly assumes that while in general x is well measured, occasional observations are afflicted with errors. She then views the available data as realizations of a random variable w which is a contaminated measure of x:

$$w \equiv z \cdot x + (1-z) \cdot v. \tag{A.1}$$

Here v is a random variable whose distribution is of no interest, and the unobservable binary random variable z indicates whether x or v is observed. When $z = 1$, realizations of x are observed and w is said to be error free. When $z = 0$, realizations of v are observed and w is said to be a data error. The relation between the observable distribution of w and the unobservable distribution of

x is given by

$$P(w) = \Pr(z=1) P(x|\, z=1) + \Pr(z=0) P(v|\, z=0), \tag{A.2}$$
$$P(x) = \Pr(z=1) P(x|\, z=1) + \Pr(z=0) P(x|\, z=0). \tag{A.3}$$

In order to make inference on features of $P(x)$, it is often assumed that the error probability $\Pr(z=0)$ is known, or that it can be bounded non-trivially from above.

It is now easy to show that when the error distribution in a convolution model has a mass point at zero, the convolution model can be treated as a mixture model. To see this, first let

$$w = x + \tilde{v}, \tag{A.4}$$

where $\tilde{v} = 0$ with probability $1-p$, $\tilde{v} = \varepsilon$ with probability p, and ε is a random variable with possibly unknown distribution. Then one can express the model in (A.4) as a special case of the model in (A.1) as follows:

$$w = zx + (1-z)(x+\varepsilon),$$

where $\Pr(z=0) = p$, and the contaminating random variable v which appeared in (A.1) has been replaced by $x + \varepsilon$.

When the data take values in the finite set X, the mixture model in (A.1) and the misclassification model in (1.1) can be related as follows. Starting from equation (A.2), notice that,$\forall$ $i, j \in X$,

$$\Pr(w=i|\, x=j) = \begin{cases} \Pr(z=1|\, x=j) + \Pr(z=0|\, x=j) \Pr(v=j|\, x=j, z=0) & \text{if } i=j, \\ \Pr(z=0|\, x=j) \Pr(v=i|\, x=j, z=0) & \text{if } i \neq j. \end{cases} \tag{A.5}$$

Assumptions on $P(z|\, x)$, typical of mixture models, translate immediately into assumptions for the misclassification model. This will be rigorously proved below. However, prior information on misclassification probabilities usually cannot be as easily incorporated in a mixture model.

To summarize, in the case of discrete variables with limited range, convolution models can be treated as mixture models, and mixture models can be treated as misclassification models based on equation (1.1). Equation (1.1) can therefore be used directly to draw inference on features of $P(x)$ and $P(y|\, x)$.

A.1 Assumptions on Π and Assumptions in the *Mixture Model*

When using mixture models it is usually assumed availability of a non-trivial upper bound on the probability of a data error. Hence, the following is maintained:

Assumption 5 $\Pr(z=0) \leq \lambda < 1$.

Additionally, often it is assumed that

Assumption 6 $z \perp x$.

Horowitz and Manski (1995) derive sharp bounds on $P(x)$ and on features of this distribution that respect stochastic dominance, when either Assumption 5 only, or both Assumptions 5-6 are maintained. They refer to the first case as "corrupted sampling," and to the second case as "contaminated sampling." In Section 3 I mentioned that the assumptions maintained by Horowitz and Manski (1995) imply Assumption 1 for the case of corrupted data, and Assumption 2 for the case of contaminated data. Here I derive this result rigorously.

Proposition 9 *a) Suppose that Assumption 5 holds. Then*

$$\sum_{h=1}^{J} \Pr(w=h, x=h) \geq 1-\lambda. \tag{A.6}$$

These bounds exhaust the implications of Assumption 5 on the structure of Π.
b) Suppose that Assumptions 5-6 jointly hold. Then

$$\pi_{jj} \equiv \Pr(w=j|\, x=j) = \Pr(z=1) + \Pr(z=0)\Pr(v=j|\, x=j, z=0) \geq 1-\lambda,\ \forall j \in X. \tag{A.7}$$

These bounds exhaust the implications of Assumptions 5-6 on the structure of Π.

Proof. Both for part a) and b), the proof is in two steps. First, I show that, given equation (A.1), Assumption 5 and the joint Assumptions 5-6 imply, respectively, (A.6) and (A.7). Then I show that for any Π such that:

1. there exists a column vector $\mathbf{p}^x = \begin{bmatrix} p_1^x & p_2^x & \ldots & p_J^x \end{bmatrix}'$ such that $p_j^x \geq 0$, $\forall\, j \in X$, $\sum_{j=1}^{J} p_j^x = 1$, and $\sum_{j=1}^{J} \pi_{ij} p_j^x = P_i^w$, $\forall\, i \in X$,
2. (A.6) and (A.7) are satisfied,

one can construct random variables $x \in X$, $v \in X$, and $z \in \{0,1\}$ such that

$$P_i^w = \Pr(z=1) \cdot \Pr(x=i|\, z=1) + \Pr(z=0) \cdot \Pr(v=i|\, z=0)\,,\ \forall\, i \in X,$$

with $\Pr(z=0) \leq \lambda$ both in case a) and b), and $z \perp x$ in case b).
a) Corrupted Sampling.

Step 1.

Let equation (A.1) and Assumption 5 hold. Then

$$\begin{aligned}
\sum_{h=1}^{J} \Pr\left(w=h,x=h\right) &= \sum_{h=1}^{J} \Pr\left(w=h\right|x=h) \Pr\left(x=h\right) \\
&= \sum_{h=1}^{J} \left[\Pr\left(z=1\right|x=h) + \Pr\left(z=0\right|x=h) \Pr\left(v=h\right|x=h,z=0)\right] \Pr\left(x=h\right) \\
&= \Pr\left(z=1\right) + \sum_{h=1}^{J} \Pr\left(v=h,x=h,z=0\right) \geq 1-\lambda,
\end{aligned}$$

where the first equality follows from Bayes Theorem, the second from equation (A.1), and the last inequality follows from Assumption 5 and the fact that $\sum\limits_{h=1}^{J} \Pr\left(v=h,x=h,z=0\right) \geq 0$.

Step 2.

Consider a matrix Π such that there exists a column vector $\mathbf{p}^x = \begin{bmatrix} p_1^x & p_2^x & \ldots & p_J^x \end{bmatrix}'$, with $\mathbf{p}^x \in \Delta_{J-1}$, such that $\sum_{j=1}^{J} \pi_{ij} p_j^x = P_i^w, \forall\, i \in X$, and $\boldsymbol{\pi}_j \in \Delta_{J-1}\ \forall j \in X$, and $\sum_{h=1}^{J} \pi_{hh} p_h^x \geq 1-\lambda$. Construct random variables $z \in \{0,1\}$, $x \in X$, and $v \in X$ such that:

$$\begin{cases}
\Pr\left(z=1\right) + \Pr\left(z=0\right) = 1, \\
1-\lambda \leq \Pr\left(z=1\right) \leq \sum\limits_{j=1}^{J} \pi_{jj} p_j^x, \\
\Pr\left(z=0,x=j\right) \geq \sum\limits_{i \neq j} \pi_{ij} p_j^x,\ \forall j \in X,
\end{cases} \tag{A.8}$$

$$\begin{cases}
\Pr\left(x=j,z=i\right) \geq 0,\ \forall j \in X,\ i=0,1, \\
\Pr\left(x=j,z=0\right) + \Pr\left(x=j,z=1\right) = p_j^x,\ \forall j \in X, \\
\sum\limits_{j=1}^{J} \Pr\left(x=j,z=i\right) = \Pr\left(z=i\right),\ i=0,1,
\end{cases} \tag{A.9}$$

and

$$\begin{cases}
\Pr\left(v=i,x=j,z=0\right) = \pi_{ij} \cdot p_j^x,\ \forall i,j \in X,\ i \neq j \\
\Pr\left(v=j,x=j,z=0\right) = \pi_{jj} p_j^x - \Pr\left(z=1,x=j\right),\ \forall i,j \in X,\ i \neq j, \\
\Pr\left(v=i,z=1\right) \geq 0,\ \forall i \in X, \\
\sum\limits_{i=1}^{J} \Pr\left(v=i,z=1\right) = \Pr\left(z=1\right).
\end{cases} \tag{A.10}$$

Notice that, given the first two equations in (A.8),

$$\lambda \geq \Pr\left(z=0\right) \geq 1 - \sum_{h=1}^{J} \pi_{hh} p_h^x = \sum_{h=1}^{J} \left\{ \sum_{i \neq h} \pi_{ih} p_h^x \right\},$$

so that the last equation in (A.8) is compatible with the previous two. Also, notice that, given (A.8), $\Pr\left(v=i,x=j,z=0\right) \in [0,1]\,, \forall\, i,j \in X$; it is straightforward to verify that $\sum_{i=1}^{J} \Pr\left(v=i,x=j,z=0\right) =$

$\Pr(x=j,z=0)$, $\forall\ j \in X$. We are now left to show that given $\mathbf{p}^x$ and equations (A.8)-(A.10), the implied $P(v)$ is a valid probability measure, and $P(v, z=0)$ is such that $P_i^w = \Pr(x=i,z=1) + \Pr(v=i,z=0)$, $\forall\ i \in X$. First, notice that

$$\begin{aligned}
\sum_{i=1}^{J} \Pr(v=i,z=0) &= \sum_{i=1}^{J} \left\{ \sum_{j=1}^{J} \Pr(v=i,x=j,z=0) \right\} \\
&= \sum_{i=1}^{J} \left\{ \sum_{j\neq i} \pi_{ij}\cdot p_j^x + \pi_{ii} p_i^x - \Pr(z=1,x=i) \right\}, \\
&= \sum_{i=1}^{J} P_i^w - \Pr(z=1) = 1 - \Pr(z=1) = \Pr(z=0),
\end{aligned}$$

so that $\sum_{i=1}^{J} \Pr(v=i,z=0) + \sum_{i=1}^{J} \Pr(v=i,z=1) = 1$. Hence, the implied $P(v)$ is a valid probability measure. Now, consider

$$\begin{aligned}
\Pr(x=i,z=1) + \Pr(v=i,z=0) &= \Pr(x=i,z=1) + \sum_{j\neq i} \pi_{ij}\cdot p_j^x + \pi_{ii} p_i^x - \Pr(z=1,x=i) \\
&= \pi_{ii}\cdot p_i^x + \sum_{j\neq i} \pi_{ij}\cdot p_j^x = P_i^w, \forall i \in X.
\end{aligned}$$

Hence, the suggested distributions of x, v, and z can be used to construct a mixture model as the one in (A.1), such that the observed vector $[\Pr(w=i), i\in X]$ is a mixture of $[\Pr(x=i|\, z=1), i \in X]$ and $[\Pr(v=i|\, z=0), i\in X]$, and $\Pr(z=0) \leq \lambda$.

b) Contaminated Sampling.

Step 1.

Let equation (A.1) and Assumptions 5-6 jointly hold. Then

$$\Pr(w=h|\, x=h) = \Pr(z=1) + \Pr(z=0)\Pr(v=h|\, x=h, z=0) \geq 1-\lambda,\ \forall\ h \in X,$$

where the first equality follows from the Law of Total Probability, Assumption 6, and equation (A.1), and the last inequality follows from Assumption 5 and the fact that $\Pr(v=h|\, x=h,z=0) \geq 0$, $\forall\ h \in X$.

Step 2.

Consider a matrix Π such that $\boldsymbol{\pi}_j \in \Delta_{J-1}$ and $\pi_{jj} \geq 1-\lambda\ \forall\ j\in X$, and for which there exists a column vector $\mathbf{p}^x = \begin{bmatrix} p_1^x & p_2^x & \dots & p_J^x \end{bmatrix}'$ such that $\mathbf{p}^x \in \Delta_{J-1}$, and $\sum_{j=1}^{J} \pi_{ij} p_j^x = P_i^w$, $\forall\ i \in X$. Choose a random variable x such that:

$$\Pr(x=j|\, z=0) = \Pr(x=j|\, z=1) = p_j^x,\ \forall j \in X.$$

Construct random variables $z \in \{0,1\}$ and $v \in X$ such that:

$$1-\lambda \leq \Pr(z=1) \leq \pi_{jj},\ \forall j\in X, \tag{A.11}$$

and, for any $j \in X$ such that $p_j^x > 0$,

$$\Pr\left(v = i \mid x = j, z = 0\right) = \frac{\pi_{ij}}{\Pr\left(z = 0\right)},\ \forall\ i \in X,\ i \neq j, \tag{A.12}$$

$$\Pr\left(v = j \mid x = j, z = 0\right) = \frac{\pi_{jj} - \Pr\left(z = 1\right)}{\Pr\left(z = 0\right)}. \tag{A.13}$$

Notice that, for any $j \in X$ such that $p_j^x > 0$, given (A.11), $\Pr\left(v = i \mid x = j, z = 0\right) \in [0, 1],\ \forall\ i \in X$; it is straightforward to verify that $\sum_{i=1}^{J} \Pr\left(v = i \mid x = j, z = 0\right) = 1,\ \forall\ j \in X$ such that $p_j^x > 0$. We are now left to show that given $\mathbf{p}^x$ and equations (A.12)-(A.13), the implied $P\left(v \mid z = 0\right)$ is a valid probability measure, and is such that $P_i^w = \Pr\left(z = 1\right) \cdot p_i^x + \Pr\left(z = 0\right) \cdot \Pr\left(v = i \mid z = 0\right),\ \forall\ i \in X$. First, notice that

$$\begin{aligned}
\textstyle\sum_{i=1}^{J} \Pr\left(v = i \mid z = 0\right) &= \textstyle\sum_{i=1}^{J} \left\{\sum_{j=1}^{J} \Pr\left(v = i \mid x = j, z = 0\right) \cdot p_j^x\right\} \\
&= \textstyle\sum_{i=1}^{J} \dfrac{\pi_{ii} - \Pr\left(z = 1\right)}{\Pr\left(z = 0\right)} \cdot p_i^x + \sum_{i=1}^{J} \left\{\sum_{j \neq i} \dfrac{\pi_{ij}}{\Pr\left(z = 0\right)} \cdot p_j^x\right\} \\
&= \textstyle\sum_{i=1}^{J} \dfrac{\pi_{ii} \cdot p_i^x + \sum_{j \neq i} \pi_{ij} \cdot p_j^x}{\Pr\left(z = 0\right)} - \dfrac{\Pr\left(z = 1\right)}{\Pr\left(z = 0\right)} \\
&= \dfrac{\sum_{i=1}^{J} P_i^w}{\Pr\left(z = 0\right)} - \dfrac{\Pr\left(z = 1\right)}{\Pr\left(z = 0\right)} = \dfrac{1 - \Pr\left(z = 1\right)}{\Pr\left(z = 0\right)} = 1.
\end{aligned}$$

Hence, the implied $P\left(v \mid z = 0\right)$ is a valid probability measure. Now, consider

$$\begin{aligned}
& \Pr\left(z = 1\right) \cdot p_i^x + \Pr\left(z = 0\right) \cdot \Pr\left(v = i \mid z = 0\right) \\
= \ & \Pr\left(z = 1\right) \cdot p_i^x + \Pr\left(z = 0\right) \cdot \left(\frac{\pi_{ii} - \Pr\left(z = 1\right)}{\Pr\left(z = 0\right)} \cdot p_i^x + \textstyle\sum_{j \neq i} \dfrac{\pi_{ij}}{\Pr\left(z = 0\right)} \cdot p_j^x\right) \\
= \ & \Pr\left(z = 1\right) \cdot p_i^x + \left(\pi_{ii} - \Pr\left(z = 1\right)\right) \cdot p_i^x + \textstyle\sum_{j \neq i} \pi_{ij} \cdot p_j^x \\
= \ & \pi_{ii} \cdot p_i^x + \textstyle\sum_{j \neq i} \pi_{ij} \cdot p_j^x = P_i^w, \forall i \in X.
\end{aligned}$$

Hence, the suggested distributions of x, v, and z can be used to construct a mixture model as the one in (A.1), such that the observed vector $\left[\Pr\left(w = i\right), i \in X\right]$ is a mixture of $\left[p_i^x, i \in X\right]$ and $\left[\Pr\left(v = i \mid z = 0\right), i \in X\right]$, $x \perp z$, and $\Pr\left(z = 0\right) \leq \lambda$. ■

Notice that (A.6) and (A.7), respectively, correspond to Assumption 1 and Assumption 2, and imply the same sets $H^{E,1}\left[\Pi^{\star}\right]$ and $H^{E,2}\left[\Pi^{\star}\right]$.

We are now ready to relate Assumptions 3 and 4 to corresponding assumptions for the mixture model in (A.1). In all that follows, suppose that Assumptions 5-6 jointly hold.

Consider first Assumption 3, and suppose that in the mixture model one maintains the following:

Assumption 7 $\Pr\left(v = j \mid x = j, z = 0\right) = k\ \forall j \in \tilde{X} \subseteq X,\ k \in [0, 1]$.

Then

$$\pi \equiv \Pr\left(w=j|\, x=j\right) = \Pr\left(z=1\right) + k\Pr\left(z=0\right) \geq 1-\lambda,\ \forall j \in \tilde{X} \subseteq X,$$

which coincides with Assumption 3.

Consider now Assumption 4, and suppose that in the mixture model one maintains the following:

Assumption 8 $\Pr\left(v=j|\, x=j, z=0\right) \geq \Pr\left(v=j+1|\, x=j+1, z=0\right), \forall j \in X.$

Then, by Assumption 8, $\forall j \in X \backslash \left\{J\right\},$

$$\begin{aligned} \Pr\left(w=j|\, x=j\right) &= \Pr\left(z=1\right) + \Pr\left(z=0\right)\Pr\left(v=j|\, x=j, z=0\right) \\ &\geq \Pr\left(z=1\right) + \Pr\left(z=0\right)\Pr\left(v=j+1|\, x=j+1, z=0\right) \\ &= \Pr\left(w=j+1|\, x=j+1\right), \end{aligned}$$

and, by Assumptions 5-6, $\Pr\left(w=J|\, x=J\right) \geq 1-\lambda$.This coincides with Assumption 4.

A.2 *Mixture Model* and Dichotomous Variables

Errors in dichotomous variables are often thought of in terms of false positives and false negatives. Part of the literature dealing with error-ridden binary data using mixture models has therefore formalized the problem as follows:

$$w = z \cdot x + \left(1-z\right)\left(1-x\right). \tag{A.14}$$

(See, for example, Copas (1988) and Horowitz and Manski (1997).) Compare (A.14) with (A.1): while in (A.1) the contaminating variable is the unknown v, in (A.14) it is implicitly assumed that $v = 1-x$. Hence, when $z=0$ the realization of w is exactly the opposite of the true realization of x, while for a general mixture model it might still be the case that when drawing from v, one draws a realization that is the same as that of x (compare with equation (A.5)). Moreover, with equation (A.14), when $z=0$ the realization of w is drawn from the distribution of $1-x$.

This difference is not necessarily a mere formalism. Suppose that the researcher believes that equation (A.14) correctly represents the relation between x and w; still, equation (A.14) does not have any content per se. However, suppose that the researcher has previous information suggesting that Assumptions 5-6 jointly hold. Then

$$\Pr\left(w=j|\, x=j\right) = \Pr\left(z=1\right) + \Pr\left(z=0\right)\Pr\left(\left(1-x\right)=j|\, x=j\right) = \Pr\left(z=1\right),\ j=0,1, \tag{A.15}$$

and Assumption 3 holds, with $\tilde{X} = X = \{0,1\}$. Hence, Proposition 7 applies, so that the identification regions $H\left[\Pr\left(x=j\right)\right],\ j=0,1$, are subsets of those generally obtained when Assumptions 5-6 are maintained with the mixture model in (A.1).

B Proofs of Propositions

B.1 Propositions in Section 2

B.1.1 Proposition 1

Proof. Let $\Pi^1 \in H^P [\Pi^\star]$. This means that $\exists\, \boldsymbol{\xi}^1 \in \Delta_{J-1}$ such that $\Pi^1 \cdot \boldsymbol{\xi}^1 = \mathbf{P}^w$. Now observe that for any $\boldsymbol{\xi} \in \Delta_{J-1}$, $\tilde{\Pi} \cdot \boldsymbol{\xi} = \mathbf{P}^w$. Hence, for any $\alpha \in (0,1)$ we have that $\left(\alpha \Pi^1 + (1-\alpha)\,\tilde{\Pi}\right) \cdot \boldsymbol{\xi}^1 = \mathbf{P}^w$, and therefore $\left(\alpha \Pi^1 + (1-\alpha)\,\tilde{\Pi}\right) \in H^P [\Pi^\star]$. To show that $H^P [\Pi^\star]$ is not star convex with respect to any other of its elements, consider a matrix $\Pi^1 \in H^P [\Pi^\star]$ with $\Pi^1 \neq \tilde{\Pi}$. Because $\Pi^1 \neq \tilde{\Pi}$, it follows that there exists an $i \in X$ such that not all elements of the $i-$th row of Π^1 are equal to P_i^w. Without loss of generality, let $i = 1$. Let $\pi_{1j}^1 > P_1^w > 0$ (a similar argument works for the case that $\pi_{1j} < P_1^w$), and without loss of generality suppose $j = 1$. Construct Π^2 as follows: $\boldsymbol{\pi}_1^2 = \mathbf{P}^w$, $\pi_{1k} = 1\ \forall\ k \in X \backslash \{1\}$. Then $\Pi^2 \in H^P [\Pi^\star]$. Let $\Pi^\alpha = \alpha \Pi^1 + (1-\alpha)\,\Pi$. Then for any $\alpha \in [0, 1 - P_1^w)$ we have that $\Pi^\alpha \notin H^P [\Pi^\star]$, because every element in the first row of the resulting matrix is strictly greater than P_1^w. ■

B.1.2 Proposition 2

The calculations which follow will show that

$$\sup_{\boldsymbol{\xi} \in \Delta_{J-1}} |Q_N(\boldsymbol{\xi}) - Q(\boldsymbol{\xi})| \xrightarrow{p} 0, \qquad \text{and} \qquad \frac{\sup_{\boldsymbol{\xi} \in \Delta_{J-1}} |Q_N(\boldsymbol{\xi}) - Q(\boldsymbol{\xi})|}{\epsilon_N} = o_p(1) .$$

The consistency result then follows from Manski and Tamer (2002), Proposition 5.

For vectors of positive probabilities $\mathbf{P}^w$ and positive constants $\boldsymbol{\mu} \equiv [\mu_1, \ldots, \mu_{\bar{q}}]$, let $Q(\boldsymbol{\xi}; \mathbf{P}^w, \boldsymbol{\mu})$ denote the value of the objective function in the nonlinear programming problem (2.8)-(2.9). As discussed in Section 2, this problem has always an optimal solution. Let $(\mathbf{v}^1, \Pi^1)$ be the maximizer for (2.8)-(2.9) when $\mathbf{P}^w = \mathbf{P}^{w,1}$ and $\boldsymbol{\mu} = \boldsymbol{\mu}^1$; we will show that a feasible vector can be constructed with $\mathbf{P}^w = \mathbf{P}^{w,2}$ and $\boldsymbol{\mu} = \boldsymbol{\mu}^2$. The strategy of this proof is similar to the one in Honore and Lleras-Muney (2004), except that here some more complications arise due to the possible nonlinearity of some of the constraints.

To simplify the notation, let $\bar{q} = q$, and assume that q_1 components of $\boldsymbol{\mu}$ are estimated for the grater-than-or-equal constraints, q_2 for the less-than-or-equal constraints, and q_3 for the equality constraints, $q_1 + q_2 + q_3 = q$. Let

$$c_1 = \min \left\{ \min_j \frac{P_j^{w,2}}{P_j^{w,1}}, \min_{l \in \{1,\ldots,q_1\}} \frac{\mu_l^2}{\mu_l^1}, \min_{m \in \{1,\ldots,q_2\}} \frac{\mu_{q_1+m}^1}{\mu_{q_1+m}^2}, \min_{s \in \{1,\ldots,q_3\}} \frac{\mu_{q_1+q_2+s}^2}{\mu_{q_1+q_2+s}^1}, \min_{s \in \{1,\ldots,q_3\}} \frac{\mu_{q_1+q_2+s}^1}{\mu_{q_1+q_2+s}^2} \right\} .$$

This implies that $0 < c_1 \le 1$. Let $\check{\Pi} \equiv c_1 \cdot \Pi^1 \ge 0$, and

$$\begin{aligned} \check{v}_j &\equiv 1 - \textstyle\sum_{i=1}^J \check{\pi}_{ij} = 1 - c_1 \sum_{i=1}^J \pi^1_{ij} \ge 1 - \sum_{i=1}^J \pi^1_{ij} \ge 0,\ j = 1, \dots, J \\ \check{v}_{J+j} &\equiv P_j^{w,2} - \textstyle\sum_{i=1}^J \check{\pi}_{ij}\xi_j = P_j^{w,2} - c_1 \sum_{i=1}^J \pi^1_{ij}\xi_j \ge \dfrac{P_j^{w,2}}{P_j^{w,1}} v^1_{J+j} \ge 0,\ j = 1, \dots, J. \end{aligned}$$

Notice that

$$\begin{aligned} \left|\check{v}_j - v^1_j\right| &\le J\left(1 - c_1\right), \\ \left|\check{v}_{J+j} - v^1_{J+j}\right| &\le (1 + J)\left(\frac{1 - c_1}{c_1}\right). \end{aligned}$$

We now turn our attention to the constraints defining $H^E\left[\Pi^\star\right]$. Suppose first that $f_l\left(\cdot\right)$, $g_m\left(\cdot\right)$ and , $h_s\left(\cdot\right)$ satisfy Assumption C1-(i). Observe that if $f_l\left(\Pi^1\right) \ge \mu^1_l$ we will have $v^1_{2J+l} = 0$; if $f_l\left(\Pi^1\right) < \mu^1_l$ we will have $v^1_{2J+l} = \mu^1_l - f_l\left(\Pi^1\right)$. For $l = 1, \dots, q_1$, let

$$\check{v}_{2J+l} \equiv \begin{cases} 0 & \text{if } f_l\left(\Pi^1\right) \ge \mu^1_l \text{ and } f_l\left(\check{\Pi}\right) \ge \mu^2_l, \\ f_l\left(\Pi^1\right) - \mu^1_l - \left(f_l\left(\check{\Pi}\right) - \mu^2_l\right) & \text{if } f_l\left(\Pi^1\right) \ge \mu^1_l \text{ and } f_l\left(\check{\Pi}\right) < \mu^2_l, \\ \mu^2_l - f_l\left(\check{\Pi}\right) & \text{if } f_l\left(\Pi^1\right) < \mu^1_l. \end{cases} \tag{B.1}$$

The suggested values of $\check{v}_{2J+l}$ are feasible. In fact, for $l = 1, \dots, q$, if $f_l\left(\Pi^1\right) \ge \mu^1_l$ the implied $\check{v}_{2J+l}$ is obviously non-negative. If $f_l\left(\Pi^1\right) < \mu^1_l$,

$$\check{v}_{2J+l} = \mu^2_l - f_l\left(\check{\Pi}\right) = \mu^2_l - f_l\left(c_1\Pi^1\right) = \tfrac{\mu^2_l}{\mu^1_l}\mu^1_l - c_1^{r_l} f_l\left(\Pi^1\right) \ge c_1^{r_l} v^1_{2J+l} \ge 0,$$

where the third equality follows from Assumption C1-(i). Moreover, by Assumption C1-(i)

$$\left|\check{v}_{2J+l} - v^1_{2J+l}\right| \le \left|\mu^2_l - \mu^1_l\right| + \left|f_l\left(\check{\Pi}\right) - f_l\left(\Pi^1\right)\right| \le M\left(\frac{1 - c_1}{c_1}\right) + \max_{\Pi \in [0,1]^{J^2}} \left|f_l\left(\Pi\right)\right| \cdot \left(1 - c_1^{r_l}\right),$$

where $\max\limits_{\Pi \in [0,1]^{J^2}} \left|f_l\left(\Pi\right)\right|$ is bounded because $f_l\left(\cdot\right)$ is a continuous function on a compact set.

Regarding the less-than-or-equal constraints, observe that under Assumption C1 a monotone transformation of $g_m\left(\Pi\right)$ and μ_{q_1+m} leaves the constraint unaltered. Hence without loss of generality when $g_m\left(\cdot\right)$ satisfies Assumptions C1-(i), we can let $r_m = 1$.

Now, notice that if $g_m\left(\Pi^1\right) \le \mu^1_{q_1+m}$ we will have $v^1_{2J+q_1+m} = 0$; if $g_m\left(\Pi^1\right) > \mu^1_{q_1+m}$ we will have $v^1_{2J+q_1+m} = g_m\left(\Pi^1\right) - \mu^1_{q_1+m}$. For $m = 1, \dots, q_2$, let

$$\check{v}_{2J+q_1+m} \equiv \begin{cases} 0 & \text{if } g_m\left(\Pi^1\right) \le \mu^1_{q_1+m} \text{ and } g_m\left(\check{\Pi}\right) \le \mu^2_{q_1+m}, \\ \mu^1_{q_1+m} - g_m\left(\Pi^1\right) + \left(\frac{1}{c_1^2} g_m\left(\check{\Pi}\right) - \mu^2_{q_1+m}\right) & \text{if } g_m\left(\Pi^1\right) \le \mu^1_{q_1+m} \text{ and } g_m\left(\check{\Pi}\right) > \mu^2_{q_1+m}, \\ \frac{1}{c_1^2} g_m\left(\check{\Pi}\right) - \mu^2_{q_1+m} & \text{if } g_m\left(\Pi^1\right) > \mu^1_{q_1+m}. \end{cases}$$

This choice of $\check{v}_{2J+q_1+m}$ satisfies the constraint in (2.9). In fact, if $g_m\left(\check{\Pi}\right) \leq \mu^2_{q_1+m}$ the constraint is satisfied with $\check{v}_{2J+q_1+m} = 0$, and in the other cases

$$\mu^2_{q_1+m} - g_m\left(\check{\Pi}\right) + \check{v}_{2J+q_1+m} \geq \mu^2_{q_1+m} - g_m\left(\check{\Pi}\right) + \left(\frac{1}{c_1^2} g_m\left(\check{\Pi}\right) - \mu^2_{q_1+m}\right) = \left(\frac{1}{c_1^2} - 1\right) g_m\left(\check{\Pi}\right) \geq 0,$$

where the last inequality follows because by Assumption C1 $g_m(\cdot)$ is non-negative on $[0,1]^{J^2}$ and $0 < c_1 \leq 1$ by construction. Notice also that the suggested values of $\check{v}_{2J+q_1+m}$ are feasible. In fact, for $m = 1, \ldots, q_2$, if $g_m\left(\Pi^1\right) \leq \mu^1_{q_1+m}$ the implied $\check{v}_{2J+q_1+m}$ is obviously non-negative, because $\frac{1}{c_1^2} g_m\left(\check{\Pi}\right) \geq g_m\left(\check{\Pi}\right)$. On the other hand, recalling that by construction $c_1 \leq \min\limits_{m \in \{1,\ldots,q_2\}} \frac{\mu^1_{q_1+m}}{\mu^2_{q_1+m}}$, if $g_m\left(\Pi^1\right) > \mu^1_{q_1+m}$,

$$\check{v}_{2J+q_1+m} = \frac{1}{c_1^2} g_m\left(\check{\Pi}\right) - \mu^2_{q_1+m} = \frac{1}{c_1^2} g_m\left(c_1 \Pi^1\right) - \mu^2_{q_1+m} = \frac{1}{c_1} g_m\left(\Pi^1\right) - \mu^2_{q_1+m} \geq \frac{1}{c_1} v^1_{2J+q_1+m} \geq 0$$

Moreover, by Assumption C1-(i)

$$\left|\check{v}_{2J+q_1+m} - v^1_{2J+q_1+m}\right| \leq \left|\mu^2_{q_1+m} - \mu^1_{q_1+m}\right| + \left|\frac{1}{c_1^2} g_m\left(\check{\Pi}\right) - g_m\left(\Pi^1\right)\right| \leq \left(\frac{1-c_1}{c_1}\right)\left(M + \max_{\Pi \in [0,1]^{J^2}} g_m(\Pi)\right),$$

where $\max\limits_{\Pi \in [0,1]^{J^2}} g_m(\Pi)$ is bounded because $g_m(\cdot)$ is a continuous function on a compact set.

Suppose now that $f_l(\cdot)$, $g_m(\cdot)$ and , $h_s(\cdot)$ satisfy Assumption C1-(ii). Then letting $\check{v}_{2J+l}$ be defined as in (B.1) and defining

$$\check{v}_{2J+q_1+m} \equiv \begin{cases} 0 & \text{if } g_m\left(\Pi^1\right) \leq \mu^1_{q_1+m} \text{ and } g_m\left(\check{\Pi}\right) \leq \mu^2_{q_1+m}, \\ \mu^1_{q_1+m} - g_m\left(\Pi^1\right) + \left(\frac{1}{c_1^{t+1}} g_m\left(\check{\Pi}\right) - \mu^2_{q_1+m}\right) & \text{if } g_m\left(\Pi^1\right) \leq \mu^1_{q_1+m} \text{ and } g_m\left(\check{\Pi}\right) > \mu^2_{q_1+m}, \\ \frac{1}{c_1^{t+1}} g_m\left(\check{\Pi}\right) - \mu^2_{q_1+m} & \text{if } g_m\left(\Pi^1\right) > \mu^1_{q_1+m}. \end{cases}$$

where t is the degree of the polynomial, one can repeat similar calculations as above, showing that these choices of $\check{v}_{2J+l}$ and $\check{v}_{2J+q_1+m}$ are feasible, satisfy the constraints in (2.9), and are such that

$$\begin{aligned} \left|\check{v}_{2J+l} - v^1_{2J+l}\right| &\leq \left(\frac{1-c_1}{c_1}\right) M + const \cdot \textstyle\sum_{j=1}^t \left(1 - c_1^j\right) \\ \left|\check{v}_{2J+q_1+m} - v^1_{2J+q_1+m}\right| &\leq \left(\frac{1-c_1}{c_1}\right) M + const \cdot \textstyle\sum_{j=1}^t \left(\frac{1-c_1^j}{c_1^j}\right) \end{aligned}$$

Finally, observe that for the equality constraints the same calculations as above can be applied to $h_k(\Pi) \geq \mu_{q_1+q_2+k}$ and $h_k(\Pi) \leq \mu_{q_1+q_2+k}$, $k = 1, \ldots, q_3$.

Hence, letting $r = \max\left(t, \max\limits_l r_l\right)$ for each $\boldsymbol{\xi}$, $Q\left(\boldsymbol{\xi}; \mathbf{P}^{w,2}, \boldsymbol{\mu}^2\right) \geq Q\left(\boldsymbol{\xi}; \mathbf{P}^{w,1}, \boldsymbol{\mu}^1\right) - const \cdot (1 - c_1)$ $- const \cdot \left(1 - c_1^r\right) - const \cdot \left(\frac{1-c_1}{c_1}\right) - const \cdot \left(\frac{1-c_1^r}{c_1^r}\right)$. Interchanging the role of $\mathbf{P}^{w,1}$ and $\mathbf{P}^{w,2}$ we get

$Q\left(\boldsymbol{\xi};\mathbf{P}^{w,1},\boldsymbol{\mu}^1\right) \geq Q\left(\boldsymbol{\xi};\mathbf{P}^{w,2},\boldsymbol{\mu}^2\right) - const\cdot(1-c_2) - const\cdot(1-c_2^r) - const\cdot\left(\frac{1-c_2}{c_2}\right) - const\cdot\left(\frac{1-c_2^r}{c_2^r}\right)$,
where

$$c_2 = \min\left\{\min_j \frac{P_j^{w,1}}{P_j^{w,2}}, \min_{l\in\{1,\dots,q_1\}} \frac{\mu_l^1}{\mu_l^2}, \min_{m\in\{1,\dots,q_2\}} \frac{\mu_{q_1+m}^2}{\mu_{q_1+m}^1}, \min_{s\in\{1,\dots,q_3\}} \frac{\mu_{q_1+q_2+s}^2}{\mu_{q_1+q_2+s}^1}, \min_{s\in\{1,\dots,q_3\}} \frac{\mu_{q_1+q_2+s}^1}{\mu_{q_1+q_2+s}^2}\right\}$$

with $0 < c_2 \leq 1$, so that

$$\left|Q\left(\boldsymbol{\xi};\mathbf{P}^{w,2},\boldsymbol{\mu}^2\right) - Q\left(\boldsymbol{\xi};\mathbf{P}^{w,1},\boldsymbol{\mu}^1\right)\right| \leq const\cdot(1-c_1) + const\cdot(1-c_2) + const\cdot(1-c_1^r) +$$
$$+ const\cdot\left(\frac{1-c_1^r}{c_1^r}\right) + const\,(1-c_2^r) + const\cdot\left(\frac{1-c_1}{c_1}\right) + const\cdot\left(\frac{1-c_2}{c_2}\right) + const\cdot\left(\frac{1-c_2^r}{c_2^r}\right).$$

Finally, under Assumption C2 the estimators $\mathbf{P}_N^w$ (as defined in (2.6)) and $\mu_{l.n}$ are root-N consistent and asymptotically normal, so that $\sup_{\boldsymbol{\xi}\in\Delta_{J-1}} |Q_N(\boldsymbol{\xi}) - Q(\boldsymbol{\xi})| = O_p\left(N^{-\frac{1}{2}}\right)$.

■

B.2 Propositions in Section 3

I first introduce and prove a Lemma that will be useful for the proof of some of the following Propositions.

Lemma 1 *Suppose that Assumption 2 holds, and that* $P_j^w > \lambda$, $j \in X$. *Then* $\frac{P_j^w-\lambda}{1-\lambda}$ *is an admissible value of* p_j^x, *and therefore solves the* $j-$*th equation of system (1.1), if and only if the following conditions jointly hold: (a)* $\pi_{jj} = 1$, *and (b) either* $\pi_{ji} = \lambda$ *or* $p_i^x = 0$, $\forall\ i \in X\backslash\{j\}$, *and* $\sum_{i\neq j}\pi_{ji}p_i^x = \lambda\frac{1-P_j^w}{1-\lambda}$. □

Proof. For $\frac{P_j^w-\lambda}{1-\lambda} > 0$ to be an admissible value of p_j^x, the $j-$th equation of system (1.1) requires that

$$\pi_{jj}\frac{P_j^w-\lambda}{1-\lambda} + \sum_{i\neq j}\pi_{ji}p_i^x = P_j^w, \tag{B.2}$$

and $\sum_{i\neq j} p_i^x = \frac{1-P_j^w}{1-\lambda}$. By Assumption 2, $\pi_{ji} \in [0,\lambda]$, $\forall\ i \in X\backslash\{j\}$ and $\pi_{jj} \in [1-\lambda, 1]$. Notice that it is possible for $\pi_{ji} = \lambda$, $\forall\ i \in X\backslash\{j\}$, because the π_{ji} are not related across i. (Recall that $1-\pi_{kk} = \sum_{l\neq k}\pi_{lk} \leq \lambda$, $\forall\ k \in X$.) Therefore,

$$\begin{aligned}\pi_{jj}\frac{P_j^w-\lambda}{1-\lambda} + \sum_{i\neq j}\pi_{ji}p_i^x &\leq \pi_{jj}\frac{P_j^w-\lambda}{1-\lambda} + \lambda\sum_{i\neq j}p_i^x \\ &= \pi_{jj}\frac{P_j^w-\lambda}{1-\lambda} + \lambda\frac{1-P_j^w}{1-\lambda} \leq \frac{P_j^w-\lambda}{1-\lambda} + \lambda\frac{1-P_j^w}{1-\lambda} = P_j^w\end{aligned}$$

Hence, equation (B.2) can be satisfied if and only if $\pi_{jj} = 1$, and $\sum_{i\neq j}\pi_{ji}p_i^x = \lambda\frac{1-P_j^w}{1-\lambda}$. For the last equality to hold, we need that either $\pi_{ji} = \lambda$, $\forall\ i \in X\backslash\{j\}$, or that for any i such that $\pi_{ji} < \lambda$, $p_i^x = 0$. Notice that we must have at least one value of $p_i^x > 0$, because $p_j^x = \frac{P_j^w-\lambda}{1-\lambda} < 1$. ■

B.2.1 Proposition 3

Proof

a) Assumption 1 holds.

Given Assumption 1, we can define $H^1[\Pi^\star]$ as follows:

$$H^1[\Pi^\star] = \left\{ \begin{array}{ll} \Pi : \boldsymbol{\pi}_j \in \Delta_{J-1} \text{ and } p_j^x \geq 0 \; \forall \; j \in X, & \\ \mathbf{P}^w \in conv\{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_J\}, \text{ and } \sum_{h=1}^J \pi_{hh} p_h^x \geq 1-\lambda & \end{array} \right\}.$$

Without loss of generality, suppose that we are interested in characterizing the identification region $H[\Pr(x=1)]$. For the first equation of system (1.1) to be satisfied we need

$$p_1^x = P_1^w - \sum_{j=2}^J \pi_{1j} p_j^x + p_1^x \sum_{i=2}^J \pi_{i1}$$

From the definition of $H^1[\Pi^\star]$ we know that

$$\lambda \geq 1 - \sum_{h=1}^J \pi_{hh} p_h^x = \sum_{h=1}^J \left\{ \sum_{i \neq h} \pi_{ih} p_h^x \right\} \geq \sum_{j=2}^J \pi_{1j} p_j^x + p_1^x \sum_{i=2}^J \pi_{i1}$$

Hence from the first equation of system (1.1) we can learn that $p_1^x \geq \max\{P_1^w - \lambda, 0\}$, and $p_1^x \leq \min\{1, P_1^w + \lambda\}$. If $P_1^w > \lambda$, the lower bound is achieved for $\sum_{j=2}^J \pi_{1j} p_j^x = \lambda$ and $\sum_{i=2}^J \pi_{i1} p_1^x = 0$. If $P_1^w < 1-\lambda$, the upper bound is achieved for $\sum_{j=2}^J \pi_{1j} p_j^x = 0$ and $\sum_{i=2}^J \pi_{i1} p_1^x = \lambda$. We are now left to show that we can find values of $p_j^x \in X \backslash \{1\}$ and $\Pi \in H^1[\Pi^\star]$ such that the corresponding $\mathbf{p}^x \in H[P(x)]$.

a.1.1) Upper Bound, with $P_1^w < 1-\lambda$.

Let $\pi_{11} = \frac{P_1^w}{(P_1^w+\lambda)}$, $\pi_{jj} = 1$, $j \in X \backslash \{1\}$, $\pi_{ij} = 0$, $i,j \in X \backslash \{1\}$, $i \neq j$, and define π_{i1}, $i \in X \backslash \{1\}$, as follows:

$$\text{if } \exists \; j > 1 : P_j^w \geq \lambda, \qquad \pi_{i1} = \begin{cases} \frac{\lambda}{(P_1^w+\lambda)} & \text{for } i = j = \min\{k = 2, \ldots, J : P_k^w \geq \lambda\} \\ 0, & \forall \; i \in X, \; i \neq \{1, j\}. \end{cases}$$

$$\text{if } P_j^w < \lambda, \forall j \in X \backslash \{1\}, \quad \pi_{i1} = \begin{cases} \frac{P_2^w}{(P_1^w+\lambda)} & \text{for } i = 2, \\ \min\left\{ \frac{\lambda}{(P_1^w+\lambda)} - \sum_{k=2}^{i-1} \frac{P_k^w}{(P_1^w+\lambda)}, \frac{P_i^w}{(P_1^w+\lambda)} \right\} & \begin{cases} \text{for } i \in X \backslash \{1,2\}, \\ i : \sum_{k=2}^{i-1} \frac{P_k^w}{(P_1^w+\lambda)} \leq \lambda, \end{cases} \\ 0 & \begin{cases} \text{for } i \in X \backslash \{1,2\}, \\ i : \sum_{k=2}^{i-1} \frac{P_k^w}{(P_1^w+\lambda)} > \lambda. \end{cases} \end{cases}$$

It is easy to show that the suggested Π belongs to $H^1[\Pi^\star]$, and allows for $p_1^x = P_1^w + \lambda$ and the implied p_j^x, $j \in X \backslash \{1\}$ to solve system (1.1). Hence, $p_1^x = P_1^w + \lambda$ is a feasible value of $\Pr(x=1)$

given the maintained assumptions. To show that $p_1^x = P_1^w + \lambda$ is the sharp upper bound on $\Pr(x=1)$, take any $\varepsilon > 0$, and let $p_1^x = P_1^w + \lambda + \varepsilon$. Then, using again the first equation of system (1.1), we have

$$P_1^w + \lambda + \varepsilon = P_1^w - \textstyle\sum_{j=2}^J \pi_{1j} p_j^x + \sum_{i=2}^J \pi_{i1} p_1^x,$$

but the right hand side of the above expression is necessarily less than or equal to $P_1^w + \lambda$. This immediately shows that there exists no value of $\Pi \in H^1[\Pi^\star]$ for which $p_1^x = P_1^w + \lambda + \varepsilon$ solves system (1.1), and therefore it is not a feasible value of $\Pr(x=1)$.

a.1.2) *Upper Bound, with* $P_1^w \geq 1 - \lambda$.
In this case the upper bound is not informative, but just set equal to 1. Let $p_1^x = 1$; this in turn implies $p_j^x = 0$, $\forall\ j \in X \backslash \{1\}$. Let $\sum_{i=2}^J \pi_{i1} = 1 - P_1^w \leq \lambda$, and $\pi_{i1} p_1^x = \pi_{i1} = P_i^w \leq \lambda$, $\forall\ i \in X$, $i \neq 1$. It is straightforward to verify that the suggested $\Pi \in H^1[\Pi^\star]$, and allows for $p_1^x = 1$, and the implied $p_j^x = 0$, $\forall\ j \in X\backslash\{1\}$, to solve system (1.1). Hence $p_1^x = 1$ is a feasible value of $\Pr(x=1)$ given the maintained assumptions.

a.2.1) *Lower Bound, with* $P_1^w > \lambda$.
Let $p_2^x = P_2^w + \lambda$, and $\pi_{12} = \frac{\lambda}{p_2^x}$, $\pi_{22} = 1 - \frac{\lambda}{p_2^x}$, and $\pi_{jj} = 1$, $\forall\ j \in X\backslash\{2\}$, so that $\pi_{i2} = 0$, $\forall\ i \in X\backslash\{2\}$, and $\pi_{ij} = 0$, $\forall\ i,j \in X$, $i \neq j$, $[i\ j] \neq [1\ 2]$. Then it is straightforward to verify that the suggested $\Pi \in H^1[\Pi^\star]$, and allows for $p_1^x = P_1^w - \lambda$ and the implied p_j^x, $j \in X\backslash\{1\}$ to solve system (1.1). Hence $P_1^w - \lambda$ is a feasible value of $\Pr(x=1)$ given the maintained assumptions. To show that $p_1^x = P_1^w - \lambda$ is the sharp lower bound on $\Pr(x=j)$, take any $0 < \varepsilon \leq P_1^w - \lambda$, and let $p_1^x = P_1^w - \lambda - \varepsilon$. Then, using again the first equation of system (1.1), we have

$$P_1^w - \lambda - \varepsilon = P_1^w - \sum_{j=2}^J \pi_{1j} p_j^x + \sum_{i=2}^J \pi_{i1} p_1^x,$$

but the right hand side of the above expression is necessarily greater than or equal to $P_1^w - \lambda$. Hence, there exists no value of $\Pi \in H^1[\Pi^\star]$ for which $p_1^x = P_1^w - \lambda - \varepsilon$ solves system (1.1), and therefore it is not a feasible value of $\Pr(x=1)$.

a.2.2) *Lower Bound, with* $P_1^w \leq \lambda$.
Then the lower bound is not informative, but just set equal to 0. Let $p_1^x = 0$; this in turn implies $\sum_{j=2}^J p_j^x = 1$. Let $\pi_{12} = \pi_{13} = \ldots = \pi_{1J} = P_1^w$. Then $\sum_{j=2}^J \pi_{1j} p_j^x = P_1^w$. Moreover $\sum_{j=2}^J P_j^w$ $= 1 - P_1^w \geq 1 - \lambda$, hence $P_j^w \leq 1 - P_1^w$ for each $j \in X\backslash\{1\}$. Let $\pi_{jj} = 1 - P_1^w$, $\forall\ j \in X\backslash\{1\}$, and $\pi_{ij} = 0$, $\forall\ i,j \in X$, $i \neq j$, $i \neq 1$. Then $p_j^x = \frac{P_j^w}{1-P_1^w} \leq 1$, $j \in X\backslash\{1\}$, and $\sum_{j=2}^J p_j^x = 1$. It follows that when $P_1^w \leq \lambda$, there exist values of $\Pi \in H^1[\Pi^\star]$ for which $p_1^x = 0$ and the implied p_j^x, $j \in X\backslash\{1\}$ solve system (1.1), and hence it's a feasible value of $\Pr(x=1)$ given the maintained assumptions.

a.3) *The all interval between the extreme points is feasible.*

To prove the claim we need to distinguish four cases: (1) $\lambda \leq P_1^w \leq 1-\lambda$; (2) $P_1^w \leq \min\{\lambda, 1-\lambda\}$; (3) $P_1^w \geq \max\{\lambda, 1-\lambda\}$; (4) $1-\lambda < P_1^w < \lambda$. Here we describe in great detail the proof for case (1); the other cases can be proved using similar arguments. See Molinari (2003) for a detailed proof of all cases.

(1) Consider the case $\lambda \leq P_1^w \leq 1-\lambda$. It then follows that

$$P_1^w - \lambda \leq p_1^x \leq P_1^w + \lambda.$$

Let $p_1^x = P_1^w + (1-2\alpha)\lambda$, for any $\alpha \in (0,1)$. We want to show that we can find values of $p_j^x \in X\backslash\{1\}$ and $\Pi \in H^1[\Pi^\star]$ such that the corresponding $\mathbf{p}^x \in H[P(x)]$. We need to distinguish two sub-cases:

1. If $\alpha \leq \frac{1}{2}$, let $\pi_{11} = \frac{P_1^w}{P_1^w+(1-2\alpha)\lambda}$, $\pi_{ij} = 0$, $\forall i = 1,\ldots,J$, $j = 2,\ldots,J$. Choose π_{j1} and p_j^x, $j \in X\backslash\{1\}$, as follows:

   (a) if $\exists\, j : P_j^w \geq 1 - \frac{P_1^w}{P_1^w+(1-2\alpha)\lambda}$,

   $$\pi_{k1} = \begin{cases} 1 - \frac{P_1^w}{P_1^w+(1-2\alpha)\lambda} & \text{for } k = j = \min\{i = 2,\ldots,J : P_i^w \geq \lambda\} \\ 0, & \forall k \in X,\ k \neq \{1,j\}. \end{cases}$$

   (b) if $P_j^w < 1 - \frac{P_1^w}{P_1^w+(1-2\alpha)\lambda}, \forall j \in X\backslash\{1\}$,

   $$\pi_{k1} = \begin{cases} P_2^w & \text{for } k = 2 \\ \min\left\{1 - \frac{P_1^w}{P_1^w+(1-2\alpha)\lambda} - \sum_{i=2}^{k-1}\pi_{i1}, P_k^w\right\} & \forall k \in X\backslash\{1,2\}, \end{cases}$$

   $$p_j^x = P_j^w - \pi_{j1}\left(P_1^w + (1-2\alpha)\lambda\right).$$

2. If $\alpha > \frac{1}{2}$, let $\pi_{jj} = 1$, $\forall j \in X\backslash\{2\}$, $\pi_{22} = \frac{P_2^w}{P_2^w+(2\alpha-1)\lambda}$, $\pi_{11} = \frac{(2\alpha-1)\lambda}{P_2^w+(2\alpha-1)\lambda}$, and $p_2^x = P_2^w + (2\alpha-1)\lambda$.

b) Assumption 2 holds.

Given Assumption 2, we can define $H^2[\Pi^\star]$ as follows:

$$H^2[\Pi^\star] = \left\{\Pi : \boldsymbol{\pi}_j \in \Delta_{J-1} \text{ and } \pi_{jj} \geq 1-\lambda \text{ and } p_j^x \geq 0\ \forall\, j \in X,\ \mathbf{P}^w \in \mathit{conv}\{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_J\}\right\}.$$

Without loss of generality, suppose that we are interested in characterizing the identification region $H[\Pr(x=1)]$.

For the first equation of system (1.1) to be satisfied, we need $\pi_{11}p_1^x + \sum_{j=2}^J \pi_{1j}p_j^x = P_1^w$,where $\sum_{j=2}^J p_j^x = 1 - p_1^x$. From the definition of $H^2[\Pi^\star]$ we know that $\pi_{1j} \leq \lambda$, $\forall j \in X\backslash\{1\}$, and $\pi_{11} \geq 1-\lambda$. Let $\sum_{j=2}^J \pi_{1j}p_j^x \leq \bar{\pi}\cdot(1-p_1^x)$, where, given the above constraints, $\bar{\pi} \in [0,\lambda]$. Then

$$p_1^x = \frac{P_1^w - \bar{\pi}}{\pi_{11} - \bar{\pi}},$$

and p_1^x is well defined as long as $\pi_{11} \neq \bar{\pi}$. We now need to distinguish a few cases.

1. If $P_1^w < \min\{\lambda, 1-\lambda\}$, one can pick $\bar{\pi} = P_1^w < \lambda$, and $p_1^x = 0$ will be the lower bound. As for the upper bound, when $P_1^w < 1-\lambda \leq \pi_{11}$, by the first equation of system (1.1) $\bar{\pi} \leq P_1^w \leq \pi_{11}$, and p_1^x is decreasing in both π_{11} and $\bar{\pi}$. Hence the upper bound is achieved for $\pi_{11} = 1-\lambda$, and $\bar{\pi} = 0$, and will be given by $p_1^x = \frac{P_1^w}{1-\lambda}$.

2. If $\lambda \leq P_1^w \leq 1-\lambda$, by the first equation of system (1.1) $\bar{\pi} \leq P_1^w \leq \pi_{11}$, and p_1^x is decreasing in both π_{11} and $\bar{\pi}$. Hence the upper bound is achieved for $\pi_{11} = 1-\lambda$, and $\bar{\pi} = 0$, and will be given by $p_1^x = \frac{P_1^w}{1-\lambda}$, and the lower bound is achieved for $\pi_{11} = 1$, and $\bar{\pi} = \lambda$, and is given by $p_1^x = \frac{P_1^w - \lambda}{1-\lambda}$.

3. If $1-\lambda \leq P_1^w \leq \lambda$, pick $\bar{\pi} = P_1^w \leq \lambda$, and $p_1^x = 0$ will be the lower bound. Pick $\pi_{11} = P_1^w \geq 1-\lambda$, and $p_1^x = 1$ will be the upper bound.

4. If $P_1^w > \max\{\lambda, 1-\lambda\}$, pick $\pi_{11} = P_1^w \geq 1-\lambda$, and $p_1^x = 1$ will be the upper bound. As for the lower bound, when $P_1^w > \lambda \geq \bar{\pi}$, by the first equation of system (1.1) $\bar{\pi} \leq P_1^w \leq \pi_{11}$, and p_1^x is decreasing in both π_{11} and $\bar{\pi}$. Hence the lower bound is achieved for $\pi_{11} = 1$, and $\bar{\pi} = \lambda$, and will be given by $p_1^x = \frac{P_1^w - \lambda}{1-\lambda}$.

To summarize, from the first equation of system (1.1) we can learn that $p_1^x \geq \max\left\{\frac{P_1^w-\lambda}{1-\lambda}, 0\right\}$ and $p_1^x \leq \min\left\{1, \frac{P_1^w}{1-\lambda}\right\}$. If $P_1^w > \lambda$, the lower bound is achieved for $\pi_{11} = 1$ and $\bar{\pi} = \lambda$. If $P_1^w < 1-\lambda$, the upper bound is achieved for $\pi_{11} = 1-\lambda$ and $\bar{\pi} = 0$. We are now left to show that we can find values of $p_j^x \in X\backslash\{1\}$ and $\Pi \in H^2[\Pi^\star]$ such that for any $p_1^x \in \left[\max\left\{\frac{P_1^w-\lambda}{1-\lambda}, 0\right\}, \min\left\{1, \frac{P_1^w}{1-\lambda}\right\}\right]$ the corresponding $\mathbf{p}^x \in H[P(x)]$. We will first show that this holds for the extreme points, and then that it holds for any point in the closed interval between the lower and the upper bound.

b.1.1) *Upper Bound, with $P_1^w < 1-\lambda$.*

Let $\pi_{11} = 1-\lambda$ and $\pi_{jj} = 1, \forall\, j > 1$. Then the system reduces to

$$\begin{cases} (1-\lambda)\frac{P_1^w}{1-\lambda} = P_1^w \\ \pi_{j1}\frac{P_1^w}{1-\lambda} + p_j^x = P_j^w, \; j = 2, \ldots, J \end{cases}$$

where $\sum_{j=2}^J \pi_{j1} = \lambda$, and $\sum_{j=2}^J P_j^w > \lambda$. Choose π_{k1}, $k \in X\backslash\{1\}$, as follows:

$$\begin{array}{ll} \text{if } \exists\, j : P_j^w \geq \lambda, & \pi_{k1} = \begin{cases} \lambda & \text{for } k = j = \min\{i = 2, \ldots, J : P_i^w \geq \lambda\} \\ 0, \;\; \forall k \in X, \; k \neq \{1, j\}. \end{cases} \\ \text{if } P_j^w < \lambda, \forall j \in X\backslash\{1\}, & \pi_{k1} = \begin{cases} P_2^w & \text{for } k = 2 \\ \min\left\{\lambda - \sum_{i=2}^{k-1}\pi_{i1}, P_k^w\right\} & \forall k \in X\backslash\{1,2\}. \end{cases} \end{array} . \tag{B.3}$$

It is easy to show that the suggested Π belongs to $H^2[\Pi^\star]$, and allows for $p_1^x = \frac{P_1^w}{1-\lambda}$ and the implied p_j^x, $j \in X\backslash\{1\}$ to solve system (1.1). Hence, $p_1^x = \frac{P_1^w}{1-\lambda}$ is a feasible value of $\Pr(x=1)$ given the maintained assumptions. To show that $p_1^x = \frac{P_1^w}{1-\lambda}$ is the sharp upper bound on $\Pr(x=1)$, take any $\varepsilon > 0$, and let $p_1^x = \frac{P_1^w}{1-\lambda} + \varepsilon$. Then, we see immediately that there is no $\pi_{11} \in [1-\lambda, 1]$ for which the first equation of system (1.1) can be satisfied: even if we let $\pi_{1j} = 0$, $\forall\, j \neq 1$, $j \in X$, we would need $\pi_{11} = 1 - \frac{\lambda P_1^w + (1-\lambda)\varepsilon}{P_1^w + (1-\lambda)\varepsilon} < 1 - \lambda$ to achieve $\left(\frac{P_1^w}{1-\lambda} + \varepsilon\right)\pi_{11} = P_1^w$. Hence, there exists no value of $\Pi \in H^2[\Pi^\star]$ for which $p_1^x = \frac{P_1^w}{1-\lambda} + \varepsilon$ solves system (1.1), and therefore it is not a feasible value of $\Pr(x=1)$.

b.1.2) *Upper Bound, with* $P_1^w \geq 1 - \lambda$.

In this case the upper bound is not informative, but just set equal to 1. Let $p_1^x = 1$; this in turn implies $p_j^x = 0$, $\forall\, j \in X \backslash \{1\}$. Let $\pi_{j1} = P_j^w$, $j = 1, \ldots, J$. It is straightforward to verify that this $\Pi \in H^2[\Pi^\star]$, and obviously allows for $p_1^x = 1$ and the implied $p_j^x = 0$, $\forall\, j \in X \backslash \{1\}$, to solve system (1.1). Hence $p_1^x = 1$ is a feasible value of $\Pr(x=1)$ given the maintained assumptions.

b.2.1) *Lower Bound, with* $P_1^w > \lambda$.

Let $\pi_{j1} = 0$, $\forall\, j \in X \backslash \{1\}$, and $\pi_{12} = \ldots = \pi_{1J} = \lambda$; then the first equation of system (1.1) is satisfied, and the implied $\Pi \in H^2[\Pi^\star]$. Let $p_j^x = \frac{P_j^w}{1-\lambda} \geq 0$, $j \in X \backslash \{1\}$. It is straightforward to verify that system (1.1) is satisfied. Hence $p_1^x = \frac{P_1^w - \lambda}{1-\lambda}$ is a feasible value for $\Pr(x=1)$ given the maintained assumptions. To show that $p_1^x = \frac{P_1^w - \lambda}{1-\lambda}$ is the sharp lower bound on $\Pr(x=1)$, take any $0 < \varepsilon \leq \frac{P_1^w - \lambda}{1-\lambda}$, and let $p_1^x = \frac{P_1^w - \lambda}{1-\lambda} - \varepsilon$. Then, we see immediately that there is no $\pi_{11} \in [1-\lambda, 1]$ for which the first equation of system (1.1) can be satisfied: even if we let $\pi_{1j} = \lambda$, $\forall\, j \neq 1$, $j \in X$, we would need $\pi_{11} = \frac{P_1^w - \lambda - \lambda(1-\lambda)\varepsilon}{P_1^w - \lambda - (1-\lambda)\varepsilon} > 1$ to achieve $\left(\frac{P_1^w - \lambda}{1-\lambda} - \varepsilon\right)\pi_{11} + \lambda\left(\frac{1 - P_1^w}{1-\lambda} + \varepsilon\right) = P_1^w$. Hence, there exists no value of $\Pi \in H^2[\Pi^\star]$ for which $p_1^x = \frac{P_1^w - \lambda}{1-\lambda} - \varepsilon$ solves system (1.1), and therefore it is not a feasible value of $\Pr(x=1)$.

b.2.2) *Lower Bound, with* $P_1^w \leq \lambda$.

Let $p_1^x = 0$; this in turn implies $\sum_{j=2}^{J} p_j^x = 1$. Let $\pi_{1j} = P_1^w$ and $\pi_{jj} = 1 - P_1^w$ $\forall\, j > 1$. Then $p_j^x = \frac{P_j^w}{1-P_1^w} \geq 0$, $j \in X \backslash \{1\}$, and $\sum_{j=2}^{J} p_j^x = 1$. It follows that when $P_1^w \leq \lambda$, there exist values of $\Pi \in H^2[\Pi^\star]$ for which $p_1^x = 0$ and the implied p_j^x, $j \in X \backslash \{1\}$, solve system (1.1), and hence it's a feasible value of $\Pr(x=1)$ given the maintained assumptions.

b.3) *The all interval between the extreme points is feasible.*

To prove the claim we need to distinguish four cases: (1) $\lambda \leq P_1^w \leq 1 - \lambda$; (2) $P_1^w \leq \min\{\lambda, 1-\lambda\}$; (3) $P_1^w \geq \max\{\lambda, 1-\lambda\}$; (4) $1 - \lambda < P_1^w < \lambda$. Here we describe in great detail the proof for case (1); the other cases can be proved using similar arguments. See Molinari (2003) for a detailed proof of all cases.

(1) Consider the case $\lambda \leq P_1^w \leq 1 - \lambda$. It then follows that $\frac{P_1^w - \lambda}{1-\lambda} \leq p_1^x \leq \frac{P_1^w}{1-\lambda}$. Let $p_1^x = \frac{P_1^w - \alpha\lambda}{1-\lambda}$, for any $\alpha \in (0,1)$. We want to show that we can find values of $p_j^x \in X \backslash \{1\}$ and $\Pi \in H^2[\Pi^\star]$ such that the corresponding $\mathbf{p}^x \in H[P(x)]$. Let $\pi_{11} = 1 - \lambda(1-\alpha)$, $\pi_{1j} = \alpha\lambda$, $\forall j \in X \backslash \{1\}$, $\pi_{ij} = 0$,

$\forall\ i, j \in X\backslash\left\{1\right\}, i \neq j$. Choose π_{j1} and p^x_j, $j \in X\backslash\left\{1\right\}$, as follows:

$$\text{if } \exists\ j : P^w_j \geq \lambda\left(1-\alpha\right), \qquad \pi_{k1} = \begin{cases} \lambda\left(1-\alpha\right) & \text{for } k = j = \min\left\{i = 2, \ldots, J : P^w_i \geq \lambda\right\} \\ 0, & \forall k \in X,\ k \neq \left\{1, j\right\}. \end{cases}$$

$$\text{if } P^w_j < \lambda\left(1-\alpha\right), \forall j \in X\backslash\left\{1\right\}, \quad \pi_{k1} = \begin{cases} P^w_2 & \text{for } k = 2 \\ \min\left\{\lambda\left(1-\alpha\right) - \sum_{i=2}^{k-1}\pi_{i1}, P^w_k\right\} & \forall k \in X\backslash\left\{1,2\right\}, \end{cases}$$

$$p^x_j = \frac{1}{1-\alpha\lambda}\left(P^w_j - \pi_{j1}\cdot\frac{P^w_1 - \alpha\lambda}{1-\lambda}\right).$$

■

B.2.2 Proposition 4

Proof. (a) Suppose, without loss of generality, that $\tilde{X} = \left\{1, 2, \ldots, h\right\}$, $2 \leq h < J$, and consider $\Pr\left(x = 1\right)$. By Lemma 1, for $\frac{P^w_1 - \lambda}{1-\lambda} > 0$ to solve the first equation of system (1.1), we need $\pi_{11} = \pi = 1$, and either $\pi_{1i} = \lambda$ or $p^x_i = 0$, $\forall\ i \in X\backslash\left\{1\right\}$, with $\sum_{i=2}^{J}\pi_{1i}p^x_i = \lambda\frac{1-P^w_1}{1-\lambda}$. Since $\pi_{22} = \pi$ by assumption, and $\pi = 1$, we have that $\pi_{12} = 0$; hence, for the first equation in system (1.1) to hold, we need $p^x_2 = 0$. Consider the second equation in system (1.1): when the first equation of the system holds, the second reduces to

$$\textstyle\sum_{i=3}^{J}\pi_{2i}p^x_i = P^w_2$$

However, for each $i \in X\backslash\left\{1\right\}$, if $\pi_{1i} = \lambda$, it follows that $\pi_{2i} = 0$, since $\sum_{k \neq l}\pi_{kl} = 1 - \pi_{ll} \leq \lambda$, $\forall\ l \in X$. On the other hand, if $\pi_{1i} < \lambda$, for the first equation in system (1.1) to hold it must be the case that $p^x_i = 0$. Hence, $\sum_{i=3}^{J}\pi_{2i}p^x_i = 0$. Therefore, since $P^w_2 > 0$, the lower bound in (3.2) is not feasible for $\Pr\left(x = 1\right)$, because the second equation of system (1.1) is not satisfied. Notice now that repeating the same argument for each of equations 3 to h in system (1.1), will imply, by a symmetry argument, that $\Pr\left(x = 1\right)$ cannot achieve the lower bound in (3.2).

For $k \in \left(X - \bar{X}\right)$, $\Pr\left(x = k\right)$ can achieve the lower bound in (3.2). Consider for example $\Pr\left(x = J\right)$. Let $\pi_{JJ} = 1$, and $\pi_{Ji} = \lambda$, $\forall\ i \in X\backslash\left\{J\right\}$. Then the last equation of system (1.1) is satisfied. These values of π_{Ji}, $i \in X$, imply that $\pi = 1 - \lambda$, and that $p^x_j = \frac{P^w_j}{1-\lambda}$ for each $j \in X\backslash\left\{J\right\}$. It is obvious that the suggested $\Pi \in H^3\left[\Pi^\star\right]$, and the implied $\mathbf{p}^x_j$ solves system (1.1).
(b). Suppose that $P^w_1 \leq \lambda$, and that $p^x_1 = 0$. Then $\sum_{j=2}^{J}p^x_j = 1$, and $p^x_j \geq 0\ \forall\ j = 2, \ldots, J$. Then the proof of Proposition 3, part $b.2.2$), applies, with $\pi = 1 - P^w_1$, $\pi_{12} = \pi_{13} = \ldots = \pi_{1J} = P^w_1$, and $\pi_{ij} = 0$, $\forall i, j \in X$, $i \neq j$, $i \neq 1$. Hence, it follows that $p^x_1 = 0$ is a value consistent with Assumption 3 if $P^w_1 \leq \lambda$. ■

B.2.3 Proposition 5

Proof. (a) Suppose, without loss of generality, that $\tilde{X} = \{1, 2, \ldots, h\}$, $2 \leq h < J$, and consider $\Pr(x = 1)$. For $p_1^x = \frac{P_1^w}{1-\lambda} < 1$ to be admissible in the first equation of system (1.1), we need $\pi = 1 - \lambda$ and $\sum_{j=2}^J \pi_{1j} p_j^x = 0$. Since $\pi_{jj} = \pi$, $\forall j \in \tilde{X}$, the second equation of the system becomes:

$$\pi_{21} \frac{P_1^w}{1-\lambda} + (1-\lambda) p_2^x + \sum_{j=3}^J \pi_{2j} p_j^x = P_2^w,$$

where $\sum_{j=3}^J p_j^x = 1 - \frac{P_1^w}{1-\lambda} - p_2^x$. Let $\sum_{j=3}^J \pi_{2j} p_j^x = \bar{\pi} \cdot \left(1 - \frac{P_1^w}{1-\lambda} - p_2^x\right)$, where $\bar{\pi} \in [0, \lambda]$, since the constraints $\pi_{ij} \leq 1 - \pi \leq \lambda$, $\forall\ i \neq j \in \tilde{X}$, and $\pi_{lk} \leq \lambda$, $\forall\ l \neq k \in \left(X - \tilde{X}\right)$, allow for $\pi_{1j} = 0$ or $\pi_{1j} = \lambda$, $\forall j = 2, \ldots, J$. It follows that

$$p_2^x = \frac{P_2^w - \bar{\pi} - (\pi_{21} - \bar{\pi}) \frac{P_1^w}{1-\lambda}}{1 - \lambda - \bar{\pi}}.$$

Notice that p_2^x must lie in $\left[0, 1 - \frac{P_1^w}{1-\lambda}\right]$. We need to distinguish three cases.

1. $1 - \lambda - \bar{\pi} > 0$. Then

   $$\frac{P_2^w - \bar{\pi} - (\pi_{21} - \bar{\pi}) \frac{P_1^w}{1-\lambda}}{1 - \lambda - \bar{\pi}} \geq 0 \Longleftrightarrow \pi_{21} \leq \bar{\pi} + (P_2^w - \bar{\pi}) \frac{(1-\lambda)}{P_1^w},$$

   and we can always find values of $\pi_{21}, \bar{\pi} \in [0, \lambda]$ for which this inequality is satisfied. For $p_2^x \leq 1 - \frac{P_1^w}{1-\lambda}$ we need

   $$\frac{P_2^w - \bar{\pi} - (\pi_{21} - \bar{\pi}) \frac{P_1^w}{1-\lambda}}{1 - \lambda - \bar{\pi}} \leq 1 - \frac{P_1^w}{1-\lambda} \Longleftrightarrow \pi_{21} \geq \frac{\lambda - 1 + P_1^w + P_2^w}{P_1^w} (1-\lambda).$$

   As long as there exist values of $\pi_{21} \leq \lambda$ that satisfy the above inequality, the upper bound in (3.2) will be admissible. However,

   $$\frac{\lambda - 1 + P_1^w + P_2^w}{P_1^w} (1-\lambda) > \lambda \Longleftrightarrow P_1^w + P_2^w > (1-\lambda) + P_1^w \frac{\lambda}{1-\lambda}.$$

   Hence, we can reject the upper bound in (3.2) if

   $$P_1^w + P_2^w > (1-\lambda) + P_1^w \frac{\lambda}{1-\lambda}. \tag{B.4}$$

2. $1 - \lambda - \bar{\pi} = 0$. Then $\pi_{21} = \frac{P_1^w + P_2^w - (1-\lambda)}{P_1^w} (1-\lambda)$. Hence, we can reject the upper bound in (3.2) if condition (B.4) is satisfied.

3. $1-\lambda-\bar{\pi}<0$. Then

$$\frac{P_2^w - \bar{\pi} - (\pi_{21}-\bar{\pi})\frac{P_1^w}{1-\lambda}}{1-\lambda-\bar{\pi}} \geq 0 \Longleftrightarrow \pi_{21} \geq \bar{\pi} + (P_2^w - \bar{\pi})\frac{(1-\lambda)}{P_1^w}.$$

As long as there exist values of $\pi_{21} \leq \lambda$ that satisfy the above inequality, the upper bound in (3.2) will be admissible. However,

$$\bar{\pi} + (P_2^w - \bar{\pi})\frac{(1-\lambda)}{P_1^w} > \lambda \Longleftrightarrow P_2^w > \bar{\pi} + \frac{P_1^w(\lambda-\bar{\pi})}{1-\lambda}$$

Hence, given that by assumption $\pi_{ij} \leq \lambda$, $\forall\, i \neq j$, $i,j \in X$, we can reject the upper bound in (3.2) if $P_2^w > \lambda$. For $p_2^x \leq 1 - \frac{P_1^w}{1-\lambda}$ we need

$$\frac{P_2^w - \bar{\pi} - (\pi_{21}-\bar{\pi})\frac{P_1^w}{1-\lambda}}{1-\lambda-\bar{\pi}} \leq 1 - \frac{P_1^w}{1-\lambda} \Longleftrightarrow \pi_{21} \leq \frac{\lambda - 1 + P_1^w + P_2^w}{P_1^w}(1-\lambda)$$

As long as there exist values of $\pi_{21} \geq 0$ that satisfy the above inequality, the upper bound in (3.2) will be admissible. However,

$$\frac{\lambda - 1 + P_1^w + P_2^w}{P_1^w}(1-\lambda) < 0 \Longleftrightarrow P_1^w + P_2^w < (1-\lambda)$$

Hence, we can reject the upper bound in (3.2) if one of the following holds: (i) $P_2^w > \lambda$, or (ii) $P_1^w + P_2^w < (1-\lambda)$.

Finally, notice that

$$\begin{cases} \text{if } \lambda \leq \frac{1}{2}, & (1-\lambda-\pi_{ij}) > 0,\ \forall\, i \neq j,\ i,j \in X, \\ \text{if } \lambda > \frac{1}{2}, & \begin{cases} P_2^w > \lambda & \Longrightarrow \begin{cases} P_1^w + P_2^w > (1-\lambda) + P_1^w \frac{\lambda}{1-\lambda}, \\ P_1^w + P_2^w > (1-\lambda) \end{cases} \\ P_1^w + P_2^w < (1-\lambda) & \Longrightarrow \begin{cases} P_1^w + P_2^w < (1-\lambda) + P_1^w \frac{\lambda}{1-\lambda}, \\ P_2^w < \lambda \end{cases} \end{cases} \end{cases} .$$

When $\lambda \leq \frac{1}{2}$, condition (B.4) is necessary and sufficient to define the cases in which the upper bound in (3.2) is not feasible. When $\lambda > \frac{1}{2}$, it can still be the case that $(1-\lambda-\bar{\pi}) > 0$ (but it does not need to be). If $P_2^w > \lambda$, (B.4) is implied, and the upper bound in (3.2) is not feasible. If $P_1^w + P_2^w < (1-\lambda)$, then condition (B.4) is not satisfied, and if $(1-\lambda-\bar{\pi}) > 0$, the upper bound in (3.2) can be feasible. Hence, when $\lambda \geq \frac{1}{2}$, $P_2^w > \lambda$ is a sufficient condition for the upper bound in (3.2) to be not feasible.

Notice now that repeating the same argument for each of equations 3 to h in system (3.3), and solving each one of them, respectively, for $p_3^x, p_4^x, \ldots, p_h^x$ as we did in equation 2 for p_2^x, will imply, by a symmetry argument, that if $\lambda \leq \frac{1}{2}$, the upper bound in (3.2) can be rejected if and only if

$$P_1^w + P_j^w > (1-\lambda) + P_1^w \frac{\lambda}{1-\lambda}, \quad \text{some } j \in \tilde{X} \backslash \{1\},$$

while if $\lambda > \frac{1}{2}$, the upper bound in (3.2) can be rejected if

$$P_j^w > \lambda, \text{ some } j \in \tilde{X} \backslash \{1\}.$$

Equations $h+1$ to J in system (3.3) do not imply any additional conditions under which the upper bound in (3.2) is not feasible. Indeed, let $k \in \left(X - \tilde{X}\right)$; then

$$\pi_{21} \frac{P_1^w}{1-\lambda} + \pi_{kk} p_k^x + \sum_{j \in X \backslash \{2,k\}} \pi_{kj} p_j^x = P_k^w.$$

Let $\pi_{kk} = 1$, and, by the same argument as above, let $\sum\limits_{j \in X \backslash \{2,k\}} \pi_{kj} p_j^x = \bar{\pi} \left(1 - \frac{P_1^w}{1-\lambda} - p_k^x\right)$, where $\bar{\pi}$ must lie in $[0, \lambda]$. Then

$$p_k^x = \frac{P_k^w - \pi_{21} \frac{P_1^w}{1-\lambda} - \bar{\pi}\left(1 - \frac{P_1^w}{1-\lambda}\right)}{1 - \bar{\pi}},$$

where $1 - \bar{\pi} \geq 1 - \lambda > 0$. It is straightforward to verify that we can find values of $\pi_{21}, \bar{\pi} \in [0, \lambda]$ for which $p_k^x \in \left[0, 1 - \frac{P_1^w}{1-\lambda}\right]$. For example, if $P_k^w \leq 1 - \frac{P_1^w}{1-\lambda}$, let $\bar{\pi} = \pi_{21} = 0$, so that $p_k^x = P_k^w$. If $P_k^w > 1 - \frac{P_1^w}{1-\lambda}$ and $P_k^w > \lambda$, let $\bar{\pi} = \pi_{21} = \lambda$, so that $p_k^x = \frac{P_k^w - \lambda}{1-\lambda} \leq 1 - \frac{P_1^w}{1-\lambda}$.
(b) Suppose that $P_1^w > 1 - \lambda$, and that $p_1^x = 1$. Then $p_j^x = 0 \ \forall \ j = 2, \ldots, J$. Then pick $\pi = P_1^w$ (notice that $P_1^w > 1 - \lambda$, hence the proposed value of π is admissible), and $\pi_{j1} = P_j^w$ $\forall j = 2, 3, \ldots, J$. Since $P_1^w > 1 - \lambda$, it follows that $P_j^w < \lambda \ \forall j = 2, 3, \ldots, J$, hence the proposed values of π_{j1}, $\forall j = 2, 3, \ldots, J$, are admissible, and therefore $p_1^x = 1$ is admissible, and hence it is the upper bound. ■

B.2.4 Proposition 6

Proof.
a) Lower Bound.
Suppose, that $j > 1$, and without loss of generality consider $\Pr(x = 2)$. By Lemma 1, for $p_2^x = \frac{P_2^w - \lambda}{1-\lambda} > 0$ to solve the second equation of system (1.1), we need $\pi_{22} = 1$, and either $\pi_{2i} = \lambda$ or $p_i^x = 0$, $\forall \ i \in X \backslash \{2\}$, with $\sum_{i \neq 2} \pi_{2i} p_i^x = \lambda \frac{1 - P_2^w}{1-\lambda}$. Since $\pi_{22} \leq \pi_{11}$ by assumption, and $\pi_{22} = 1$, we have that $\pi_{11} = 1$; hence, the first equation of system (1.1) reduces to

$$\textstyle\sum_{i=3}^{J} \pi_{1i} p_i^x = P_1^w$$

However, for each $i \in X \backslash \{1, 2\}$, if $\pi_{2i} = \lambda$, it follows that $\pi_{1i} = 0$, since $\sum_{k \neq l} \pi_{kl} = 1 - \pi_{ll} \leq \lambda$, $\forall$ $l \in X$. On the other hand, if $\pi_{2i} < \lambda$, for the second equation in system (1.1) to hold it must be the case that $p_i^x = 0$. Hence, $\sum_{i=3}^{J} \pi_{1i} p_i^x = 0$. Therefore, since $P_1^w > 0$, the lower bound in (3.2) is not feasible for $\Pr(x = 2)$. Notice now that repeating the same argument for $\Pr(x = 3)$, will imply that $\Pr(x=3)$ cannot achieve the lower bound in (3.2). Similarly, $\Pr(x=j)$ cannot achieve the lower bound in (3.2).

Consider now $\Pr(x=1)$, and let $\pi_{11}=1$, and $\pi_{1i}=\lambda$, $\forall\ i \in X\backslash\{1\}$. Then the first equation of system (1.1) is satisfied. Let $p_j^x = \frac{P_j^w}{1-\lambda}$ and $\pi_{jj} = 1-\lambda$ for each $j \in X\backslash\{1\}$. It is obvious that the suggested $\Pi \in H^4[\Pi^\star]$, and the implied $\mathbf{p}_j^x$ solves system (1.1).

b) Upper Bound.
First, let $j=1$, and $P_1^w < (1-\lambda)$. Then, as shown in the proof of Proposition 5, for $p_1^x = \frac{P_1^w}{1-\lambda}$ we need $\pi_{11} = 1-\lambda$ and $\sum_{i=2}^{J} \pi_{1i} p_i^x = 0$. But by Assumption 4, $\pi_{11} \geq \pi_{22} \geq \ldots \geq \pi_{JJ} \geq 1-\lambda$, and therefore for $p_1^x = \frac{P_1^w}{1-\lambda}$ to solve the first equation of system (1.1) we need $\pi_{jj} = 1-\lambda$, $\forall\ j \in X$, and we are back to the case of constant probability of correct report, with $\tilde{X} = X$; the result of Proposition 5 part (b) applies. Now let $j>1$, and $P_j^w < (1-\lambda)$. Then, again, for $p_j^x = \frac{P_j^w}{1-\lambda}$ we need $\pi_{jj} = 1-\lambda$ and $\sum_{i \neq j} \pi_{ji} p_i^x = 0$. But by Assumption 4, $\pi_{jj} \geq \pi_{(j+1)(j+1)} \geq \ldots \geq \pi_{JJ} \geq 1-\lambda$, and therefore we need $\pi_{kk} = 1-\lambda$, $\forall\ k \in \{j, j+1, \ldots, J\}$, and we are back to the case of constant probability of correct report, with $\tilde{X} = \{j, j+1, \ldots, J\}$; the result of Proposition 5 applies. ■

B.2.5 Proposition 7

Proof. With dichotomous variables, $p_1^x(\pi) = \frac{P_1^w-(1-\pi)}{\pi-(1-\pi)} = \frac{1}{2}\left[\frac{2P_1^w-1}{2\pi-1}+1\right]$, $\pi \in H^3[\Pi^\star]$. Hence,

1. If $\lambda < \frac{1}{2}$ $P_1^w \geq \frac{1}{2}$, then $1-\pi \leq P_1^w \leq \pi$ and $\frac{\partial p_1^x(\pi)}{\partial \pi} \leq 0$. Hence the lower bound on $\Pr(x=1)$ will be achieved for $\pi=1$ and the upper bound for $\pi = \max(1-\lambda, P_1^w)$.

2. If $\lambda \geq \frac{1}{2}$ $P_1^w \geq \frac{1}{2}$, then for $p_1^x \in [0,1]$ we need one of the following: (a) $1-\pi \leq P_1^w \leq \pi \implies \pi \geq P_1^w \geq \frac{1}{2}$; or (b) $\pi \leq P_1^w \leq 1-\pi \implies \pi \leq 1-P_1^w < \frac{1}{2}$; additionally, we need $\pi \geq 1-\lambda$. Hence, the feasible values of π are given by $\pi \in [1-\lambda, 1-P_1^w] \cup [P_1^w, 1]$. Notice that if $\lambda < P_1^w$, the feasible values of π are given by $\pi \in [P_1^w, 1]$, and p_1^x is decreasing in π; therefore the lower bound is achieved for $\pi=1$ and the upper bound for $\pi = P_1^w$. When $\lambda > P_1^w$, for values of $\pi \in [P_1^w, 1]$ the previous result applies. For values of $\pi \in [1-\lambda, 1-P_1^w]$ p_1^x is decreasing in π; therefore the upper bound is achieved for $\pi = 1-\lambda$ and the lower bound for $\pi = 1-P_1^w$.

3. If $\lambda < \frac{1}{2}$ $P_1^w < \frac{1}{2}$, then $1-\pi \leq P_1^w \leq \pi$ and $\frac{\partial p_1^x(\pi)}{\partial \pi} \geq 0$. Hence the lower bound on $\Pr(x=1)$ will be achieved for $\pi = 1-\min(\lambda, P_1^w)$ and the upper bound for $\pi=1$.

4. If $\lambda \geq \frac{1}{2}$ $P_1^w < \frac{1}{2}$, then for $p_1^x \in [0,1]$ we need one of the following: (a) $1-\pi \leq P_1^w \leq \pi \implies \pi \geq 1-P_1^w > \frac{1}{2}$; or (b) $\pi \leq P_1^w \leq 1-\pi \implies \pi \leq P_1^w < \frac{1}{2}$; additionally, we need $\pi \geq 1-\lambda$. Hence, the feasible values of π are given by $\pi \in [1-\lambda, P_1^w] \cup [1-P_1^w, 1]$. Notice that if $1-\lambda > P_1^w$, the feasible values of π are given by $\pi \in [1-P_1^w, 1]$, and p_1^x is increasing

in π; therefore the lower bound is achieved for $\pi = 1 - P_1^w$ and the upper bound for $\pi = 1$. When $1 - \lambda < P_1^w$, for values of $\pi \in [1 - P_1^w, 1]$ the previous result applies. For values of $\pi \in [1 - \lambda, P_1^w]$ p_1^x is increasing in π; therefore the upper bound is achieved for $\pi = P_1^w$ and the lower bound for $\pi = 1 - \lambda$.

To show that these bounds are a subset of those in (3.2), assume that $P_1^w \geq 0.5$ (a similar argument goes for the other case). If $\lambda < \frac{1}{2}$, consider the lower bound; then $P_1^w \geq \max\left(\frac{P_1^w - \lambda}{1-\lambda}, 0\right)$. Indeed, $P_1^w > 0$; on the other hand, if $P_1^w > \lambda$, then $P_1^w - \lambda P_1^w - P_1^w + \lambda = \lambda\left(1 - P_1^w\right) \geq 0$. Consider now the upper bound; then $\min\left(\frac{P_1^w - \lambda}{1-2\lambda}, 1\right) \leq \min\left(\frac{P_1^w}{1-\lambda}, 1\right)$. Indeed, if both sides are equal to 1 the equality is trivially satisfied; hence, suppose $P_1^w < 1 - \lambda$. Then $\frac{P_1^w - \lambda}{1-2\lambda} < 1$, since $P_1^w - \lambda - (1 - 2\lambda) = P_1^w - (1 - \lambda) < 0$ (similarly, note that if $\frac{P_1^w - \lambda}{1-2\lambda} < 1$, also $\frac{P_1^w}{1-\lambda} < 1$). Compare $\frac{P_1^w - \lambda}{1-2\lambda}$ and $\frac{P_1^w}{1-\lambda}$:

$$\frac{P_1^w}{1-\lambda} - \frac{P_1^w - \lambda}{1-2\lambda} = \frac{\lambda\left(1-\lambda\right) - \lambda P_1^w}{\left(1-2\lambda\right)\left(1-\lambda\right)} > 0 \text{ if } P_1^w < 1 - \lambda$$

Hence, $\frac{P_1^w}{1-\lambda} > \frac{P_1^w - \lambda}{1-2\lambda}$ if $\frac{P_1^w}{1-\lambda} < 1$. If $\lambda \geq \frac{1}{2}$ and $P_1^w > \lambda$, the bound in (3.2) is given by $\left[\frac{P_1^w - \lambda}{1-\lambda}, 1\right]$, and the same argument as above works (in particular, $P_1^w \geq \frac{P_1^w - \lambda}{1-\lambda}$). If $\lambda \geq P_1^w \geq \frac{1}{2} \geq 1 - \lambda$, then the bound in (3.2) is given by $[0, 1]$. ■

B.2.6 Proposition 8

Proof. In this case, $p_1^x(\pi) = \frac{P_1^w - (1-\pi_{22})}{\pi_{11} - (1-\pi_{22})}$, $(\pi_{11}, \pi_{22}) \in H^4[\Pi^\star]$. Hence,

1. If $\lambda < \frac{1}{2}$, $1 - \pi_{22} \leq P_1^w \leq \pi_{11}$, and $p_1^x(\pi)$ is increasing in π_{22} and decreasing in π_{11}. Hence the lower bound is achieved for $\pi_{22} = 1 - \lambda$ and $\pi_{11} = 1$. The upper bound is achieved with $\pi_{22} = \pi_{11}$, since π_{11} bounds π_{22} from above. Hence if $P_1^w \geq \frac{1}{2}$, the upper bound is achieved for $\pi_{11} = \pi_{22} = \max\left(1 - \lambda, P_1^w\right)$. If $P_1^w < \frac{1}{2}$, the upper bound is achieved for $\pi_{11} = \pi_{22} = 1$.

2. If $\lambda \geq \frac{1}{2}$ and $P_1^w < \frac{1}{2}$, either $1 - \pi_{22} \leq P_1^w \leq \pi_{11}$ or $1 - \pi_{22} \geq P_1^w \geq \pi_{11}$. Hence, either $\pi_{11} \in [1 - P_1^w, 1]$ and $\pi_{22} \in [1 - P_1^w, \pi_{11}]$, or $\pi_{11} \in [1 - \lambda, P_1^w]$ and $\pi_{22} \in [1 - \lambda, \pi_{11}]$. In the first case case p_1^x is increasing in π_{22} and decreasing in π_{11}; the lower bound is achieved for $\pi_{11} = 1$, $\pi_{22} = 1 - P_1^w$. The upper bound is achieved with $\pi_{22} = \pi_{11} = 1$. In the second case p_1^x is decreasing in π_{22} and increasing in π_{11}; the lower bound is achieved with $\pi_{22} = \pi_{11} = 1 - \lambda$. The upper bound is achieved with $\pi_{11} = P_1^w$ and $\pi_{22} = 1 - \lambda$.

3. If $\lambda \geq \frac{1}{2}$ and $P_1^w \geq \frac{1}{2}$, consider the following two cases. If $\lambda > P_1^w$ then $\pi_{11} = \pi_{22} = 1 - P_1^w$ are admissible values, and the implied $p_1^x = 0$. Also, $\pi_{11} = P_1^w$ is an admissible value, and the implied $p_1^x = 1$. If $\lambda < P_1^w$ then $\pi_{11} \in [P_1^w, 1]$, $\pi_{22} \in [1 - \lambda, \pi_{11}]$ and $1 - \pi_{22} \leq P_1^w \leq \pi_{11}$. Then p_1^x is decreasing in π_{11} and increasing in π_{22}. Hence the lower bound is achieved for $\pi_{11} = 1$ and $\pi_{22} = 1 - \lambda$, and the upper bound is achieved with $\pi_{22} = \pi_{11} = P_1^w$. ■

References

Abrevaya, J., and J. A. Hausman (1999): "Semiparametric Estimation with Mismeasured Dependent Variables: An Application to Duration Models for Unemployment Spells," *Annales d'Economie et de Statistique*, 55-56, 243–275.

Aigner, D. J. (1973): "Regression with a Binary Independent Variable Subject to Errors of Observation," *Journal of Econometrics*, 1, 49–60.

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2003): "Changes in the Distribution of Male and Female Wages Accounting for Employment Composition," Discussion paper, Institute for Fiscal Studies.

Bollinger, C. R. (1996): "Bounding Mean Regressions When a Binary Regressor is Mismeasured," *Journal of Econometrics*, 73, 387–399.

Bound, J., C. Brown, and N. Mathiowetz (2001): "Measurement Error in Survey Data," in *Handbook of Econometrics, Vol. 5*, ed. by J. J. Heckman, and E. Leamer, pp. 3705–3843, North Holland. Elsevier Science.

Bross, I. (1954): "Misclassification in 2 X 2 Tables," *Biometrics*, 10(4), 478–486.

Campbell, S. L., and C. D. Meyer (1991): *Generalized Inverses of Linear Transformations.* Dover Publications, Inc., New York.

Card, D. (1996): "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis," *Econometrica*, 64(4), 957–979.

Chernozhukov, V., H. Hong, and E. Tamer (2004): "Parameter Set Inference in a Class of Econometric Models," Discussion paper, Northwestern University.

Copas, J. B. (1988): "Binary Regression Models for Contaminated Data," *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2), 225–265, With Comments.

Cox, D. R., and D. V. Hinkley (1974): *Theoretical Statistics.* Chapman and Hall, London, UK.

Dominitz, J., and R. P. Sherman (2003): "Identification and Estimation of Bounds on School Performance Measures: A Nonparametric Analysis of a Mixture Model with Verification," Discussion paper, CalTech.

Dustmann, C., and A. van Soest (2000): "Parametric and Semiparametric Estimation in Models with Misclassified Dependent Variables," *IZA Discussion Paper 218.*

GONG, G., A. S. WHITTEMORE, AND S. GROSSER (1990): "Censored Survival Data With Misclassified Covariates: A Case Study of Breast Cancer Mortality," *Journal of the American Statistical Association*, 85(409), 20–28.

GUSTMAN, A. L., O. S. MITCHELL, A. A. SAMWICK, AND T. L. STEINMEIER (2000): "Evaluating Pension Entitlements," in *Forecasting Retirement Needs and Retirement Wealth*, ed. by O. S. Mitchell, P. B. Hammond, and A. M. Rappaport. University of Pennsylvania.

GUSTMAN, A. L., AND T. L. STEINMEIER (2001): "What People Don't Know About Their Pension and Social Security," in *Public Policies and Private Pensions*, ed. by W. G. Gale, J. B. Shoven, and M. J. Warshawsky, Washington D.C. Brookings Institution.

HAMPEL, F. R. (1974): "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association*, 69(346), 383–393.

HAMPEL, F. R., E. M. RONCHETTI, P. J. ROUSSEEUW, AND W. A. STAHEL (1986): *Robust Statistics: The Approach Based on Influence Functions.* John Wiley and Sons.

HAUSMAN, J., J. ABREVAYA, AND F. M. SCOTT-MORTON (1998): "Misclassification of the Dependent Variable in a Discrete-Response Setting," *Journal of Econometrics*, 87, 239–269.

HONORE, B., AND A. LLERAS-MUNEY (2004): "Bounds in Competing Risks Models and the War on Cancer," Discussion paper, Princeton University.

HONORE, B. E., AND E. TAMER (2003): "Bounds on Parameters in Dynamic Discrete Choice Models," Discussion paper, Princeton University.

HORN, R. A., AND C. R. JOHNSON (1999): *Matrix Analysis.* Cambridge University Press, New York.

HOROWITZ, J. L., AND C. F. MANSKI (1995): "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63(2), 281–302.

——— (1997): "What Can Be Learned About Population Parameters When the Data are Contaminated," in *Handbook of Statistics, Vol. 15*, ed. by G. S. Maddala, and C. R. Rao, pp. 439–466, North Holland. Elsevier Science B. V.

HOROWITZ, J. L., AND C. F. MANSKI (1998): "Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations," *Journal of Econometrics*, 84, 37–58.

Horowitz, J. L., and C. F. Manski (2000): "Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95(449), 77–84.

Hotz, V. J., C. H. Mullin, and S. G. Sanders (1997): "Bounding Causal Effects Using Data from a Contaminated Natural Experiment: Analyzing the Effects of Teenage Childbearing," *Review of Economic Studies*, 64, 575–603.

Hu, Y. (2003): "Bounding Parameters in a Linear Regression Model with a Mismeasured Regressor Using Additional Information," Discussion paper, University of Texas at Austin.

Imbens, G. W., and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72(6), 1845–1857.

Kane, T. J., C. E. Rouse, and D. Staiger (1999): "Estimating Returns to Schooling When Schooling is Misreported," *NBER Working Paper 7235.*

Klepper, S. (1988): "Bounding the Effects of Measurement Error in Regressions Involving Dichotomous Variables," *Journal of Econometrics*, 37, 343–359.

Klepper, S., and E. E. Leamer (1984): "Consistent Sets of Estimates for Regressions with Errors in All Variables," *Econometrica*, 52(1), 163–183.

Kreider, B., and J. Pepper (2004): "Inferring Disability Status from Corrupt Data," Discussion paper, Iowa State University.

Lewbel, A. (2000): "Identification of the Binary Choice Model With Misclassification," *Econometric Theory*, 16, 603–609.

Mahajan, A. (2003): "Misclassified Regressors in Binary Choice Models," Discussion paper, Stanford University.

Manski, C. F. (2003): *Partial Identification of Probability Distributions*, Springer Series in Statistics. Springer-Verlag, New York.

Manski, C. F., and E. Tamer (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70(2), 519–546.

Mellow, W., and H. Sider (1983): "Accuracy of Response in Labor Market Surveys: Evidence and Implications," *Journal of Labor Economics*, 1(4), 331–344.

Molinari, F. (2003): "Contaminated, Corrupted, and Missing Data," Ph.D. thesis, Northwestern University, Available at: http://www.arts.cornell.edu/econ/fmolinari/dissertation.pdf.

MOORE, J. C., K. H. MARQUIS, AND K. BOGEN (1996): "The SIPP Cognitive Research Evaluation Experiment: Basic Results and Documentation," *Unpublished Report*, U.S. Bureau of the Census.

MUNKRES, J. R. (1991): *Analysis on Manifolds.* Addison-Wesley.

POTERBA, J. M., AND L. H. SUMMERS (1995): "Unemployment Benefits and Labor Market Transitions: A Multinomial Logit Model with Errors in Classification," *The Review of Economics and Statistics*, 77(2), 201–216.

RAMALHO, E. A. (2002): "Regression Models for Choice-Based Samples with Misclassification in the Response Variable," *Journal of Econometrics*, 106, 171–201.

RAO, C. R. (1973): *Linear Statistical Inference and its Applications.* Wiley, New York.

ROCKAFELLAR, R. T. (1970): *Convex Analysis.* Princeton University Press, Princeton, New Jersey.

Table 1: Identifying Power of Assuming Monotonicity in Correct Reporting or Constant Probability of Correct Report vs. Base-Case, with Dichotomous Variables, for Different Values of λ

	Maintained Assumptions		
	Base-Case	**Monotonicity in Correct Reporting**	**Constant Probability of Correct Report**
λ	$H\left[\Pr\left(x=1\right)\right]$	$H\left[\Pr\left(x=1\right)\right]$	$H\left[\Pr\left(x=1\right)\right]$
1.000	$[0,1]$	$[0,0.34]\cup[0.66,1]$	$[0,0.34]\cup[0.66,1]$
0.750	$[0,1]$	$[0,0.34]\cup[0.82,1]$	$[0,0.34]\cup[0.82,1]$
0.400	$[0.00,0.57]$	$[0.00,0.34]$	$[0.00,0.34]$
0.250	$[0.12,0.45]$	$[0.12,0.34]$	$[0.18,0.34]$
0.100	$[0.27,0.38]$	$[0.27,0.34]$	$[0.30,0.34]$

Table 2: Identifying Power of Assuming Monotonicity in Correct Reporting or Constant Probability of Correct Report vs. Base-Case

	Maintained Assumptions				**Exact Value**
	Base-Case	**Monotonicity in Correct Reporting**	**Constant Probability of Correct Report**		
			$\tilde{X}=\{1,2\}$	$\tilde{X}=X$	
$\Pr\left(x=1\right)$	$[0.180,0.425]$	$[0.180,0.415]$	$[0.235,0.415]$	$[0.235,0.415]$	0.3
$\Pr\left(x=2\right)$	$[0.434,0.687]$	$[0.525,0.687]$	$[0.525,0.687]$	$[0.551,0.687]$	0.6
$\Pr\left(x=3\right)$	$[0.000,0.138]$	$[0.000,0.138]$	$[0.000,0.138]$	$[0.000,0.137]$	0.1
$E\left(x\right)$	$[1.575,1.955]$	$[1.585,1.955]$	$[1.585,1.899]$	$[1.585,1.899]$	1.8

Table 3: Percentage with Self Reported Plan Type Conditional on Firm Report of Plan Type, for Respondents Reporting Pension Coverage on Current Job with a Matched Employer Plan Description. Sample Size: 2,907. Source: Gustman and Steinmeier (2001), Table 6C.

	Provider Report		
Self Report	DB	DC	Both
DB	0.56	0.26	0.45
DC	0.15	0.54	0.18
Both	0.27	0.18	0.35
Don't Know	0.02	0.02	0.02

Table 4: True Fractions of Pension Plan Types for the Subset of Respondents with Matched Data for 1992, as Calculated by Gustman and Steinmeier (2001), Table 6A, and Reported Fractions of Pension Plan Types for 1992 and 1998 (Author's Calculations).

	$t = 1992$			$t = 1992$		$t = 1998$	
	Point Est.	Bootstrap 95% C. I.		Point Est.	Bootstrap 95% C. I.	Point Est.	Bootstrap 95% C. I.
$\Pr_t(x = 1 \| s = 1)$	0.48	$[0.46, 0.50]$	$\Pr_t(w = 1)$	0.42	$[0.41, 0.44]$	0.28	$[0.25, 0.30]$
$\Pr_t(x = 2 \| s = 1)$	0.21	$[0.19, 0.22]$	$\Pr_t(w = 2)$	0.32	$[0.31, 0.33]$	0.38	$[0.35, 0.41]$
$\Pr_t(x = 3 \| s = 1)$	0.31	$[0.29, 0.33]$	$\Pr_t(w = 3)$	0.26	$[0.24, 0.27]$	0.34	$[0.31, 0.37]$
Sample Size	$n = 2,907$		Sample Size	$N = 4,354$		$N = 1,124$	

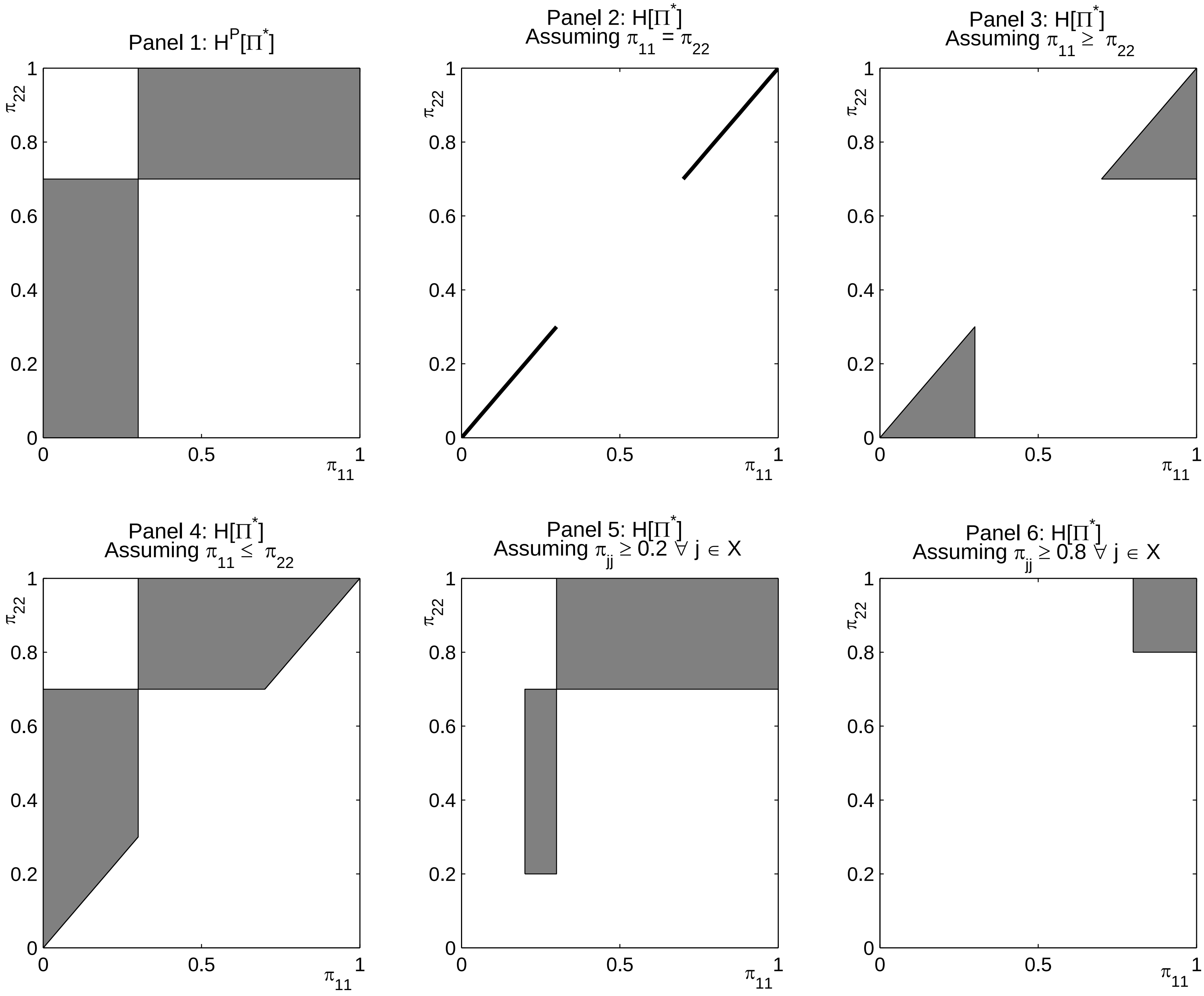
Table 5: Implications of Assumption E1 - No Selection - and Assumption E2 - No Selection and No Variation Over Time - for the Identification Regions of $[\Pr_t(x = j), j \in X]$, $t = 1992, 1998$

Maintained Assumptions:	$t = 1992$: **No Selection**		$t = 1998$: **No Selection and No Variation Over Time**	
	Point Estimate	Bootstrap 95% C. I	Point Estimate	Bootstrap 95% C. I.
$\left(\Pi^{\star 1}_{1992}\right)^{-1} \cdot \mathbf{P}^{w,t}$	0.30	$[0.27, 0.50]$	-0.86	$[-1.76, -0.42]$
	0.39	$[0.37, 0.45]$	0.48	$[0.30, 0.62]$
	0.31	$[0.07, 0.35]$	1.38	$[0.89, 2.38]$
Sample Size	$N = 4,354$		$N = 1,124$	

Table 6: Identification Regions in Cases 1-2 for $\Pr_{1998}(x = j)$, and Point Estimates for $\Pr_{1992}(x = j)$

Maintained	$H[\Pr_t(x = 1)]$		$H[\Pr_t(x = 2)]$		$H[\Pr_t(x = 3)]$	
Assumptions:	Estimate	95% CI	95% CI	95% CI	95% CI	95% CI
$t = 1992$	0.42	$[0.40, 0.50]$	0.27	$[0.25, 0.30]$	0.31	$[0.22, 0.34]$
Case 1, 1998	$[0.00, 0.42]$	$[0.00, 0.44]$	$[0.11, 0.72]$	$[0.10, 0.87]$	$[0.00, 0.89]$	$[0.00, 0.91]$
Case 2, 1998	$[0.00, 0.28]$	$[0.00, 0.34]$	$[0.35, 0.61]$	$[0.28, 0.80]$	$[0.11, 0.50]$	$[0.00, 0.67]$
Sample size	$N = 4,354$ for 1992, $N = 1,124$ for 1998					

Figure 1: Geometry of the Set $H^P[\Pi^*]$, and of the Set $H[\Pi^*]$ Under Different Assumptions, when J = 2 and Pr(w = 1) = 0.3



Panel 1: $H^P[\Pi^*]$

Panel 2: $H[\Pi^*]$ Assuming $\pi_{11} = \pi_{22}$

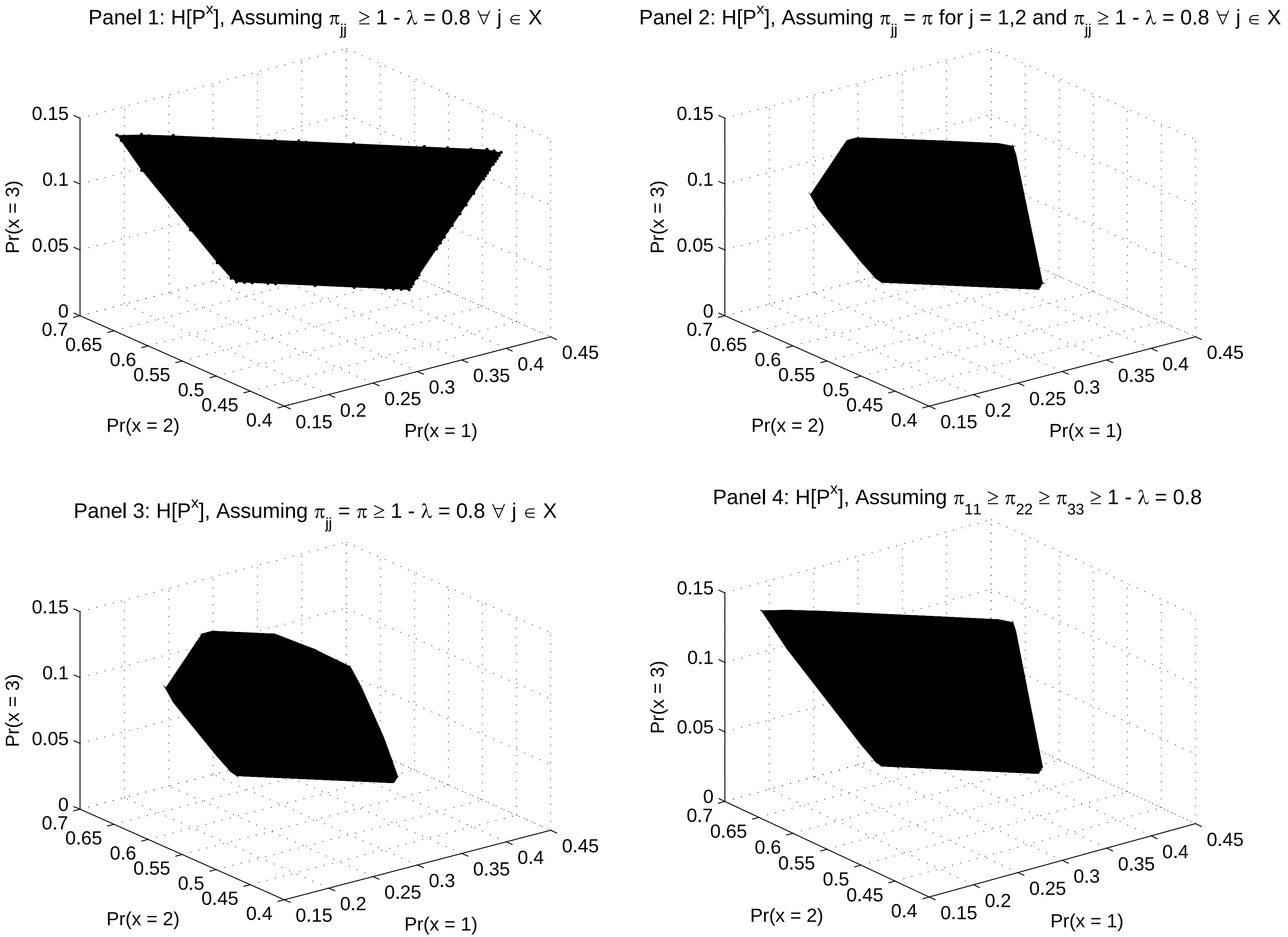
Panel 3: $H[\Pi^*]$ Assuming $\pi_{11} \geq \pi_{22}$

Panel 4: $H[\Pi^*]$ Assuming $\pi_{11} \leq \pi_{22}$

Panel 5: $H[\Pi^*]$ Assuming $\pi_{jj} \geq 0.2 \ \forall \ j \in X$

Panel 6: $H[\Pi^*]$ Assuming $\pi_{jj} \geq 0.8 \ \forall \ j \in X$

Figure 2: Comparison of the Identification Power of Different Assumptions for $H[P^x]$



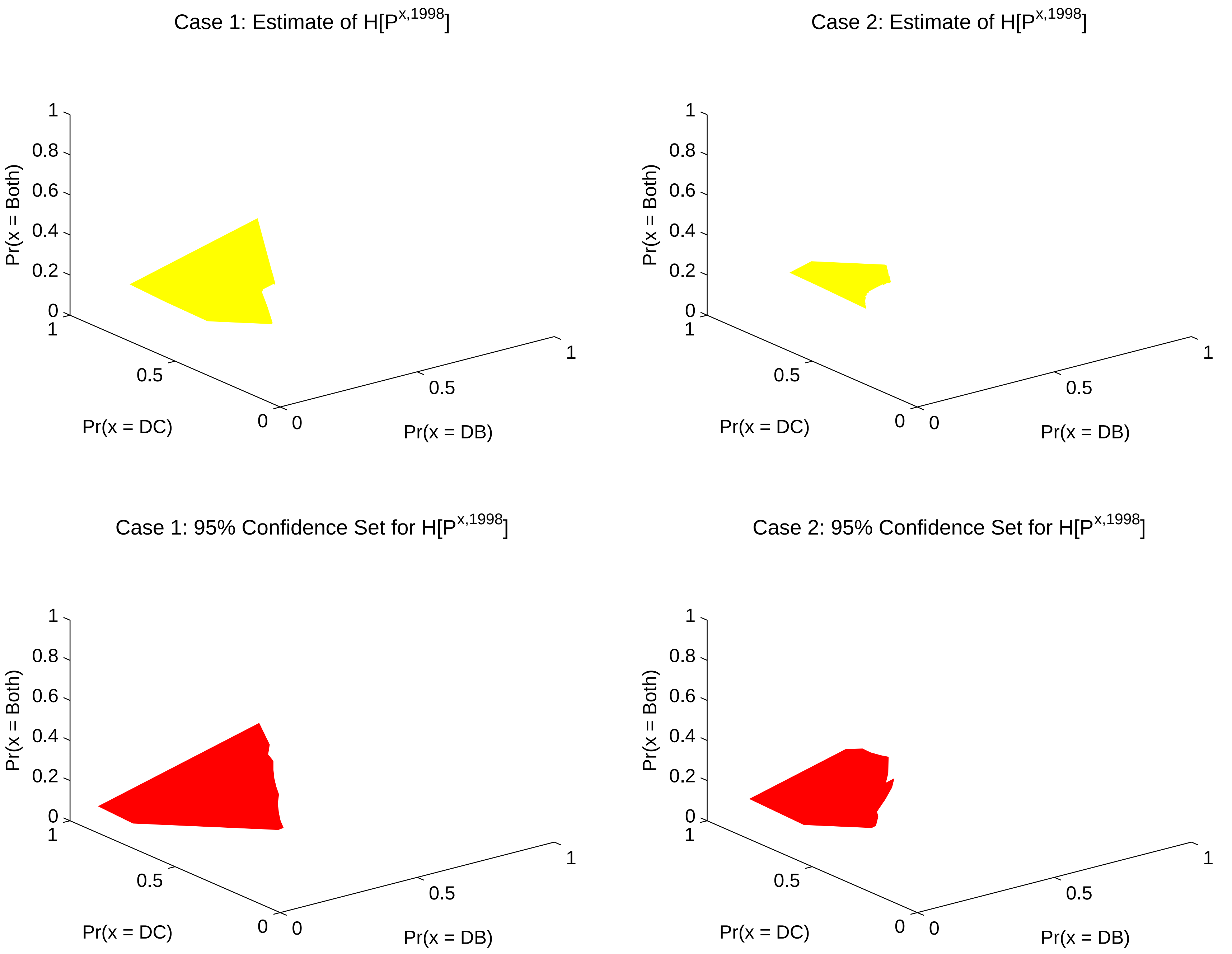
Panel 1: $H[P^x]$, Assuming $\pi_{jj} \geq 1 - \lambda = 0.8 \ \forall \ j \in X$

Panel 2: $H[P^x]$, Assuming $\pi_{jj} = \pi$ for j = 1,2 and $\pi_{jj} \geq 1 - \lambda = 0.8 \ \forall \ j \in X$

Panel 3: $H[P^x]$, Assuming $\pi_{jj} = \pi \geq 1 - \lambda = 0.8 \ \forall \ j \in X$

Panel 4: $H[P^x]$, Assuming $\pi_{11} \geq \pi_{22} \geq \pi_{33} \geq 1 - \lambda = 0.8$

Figure 3: Identification Regions and Confidence Sets for $H[P^{x,1998}]$ Under Different Assumptions



Case 1: Estimate of $H[P^{x,1998}]$

Case 2: Estimate of $H[P^{x,1998}]$

Case 1: 95% Confidence Set for $H[P^{x,1998}]$

Case 2: 95% Confidence Set for $H[P^{x,1998}]$