Institute for Policy Research
Northwestern University
Working Paper Series

WP-08-10

When Do Latent Class Models Overstate Accuracy for Binary Classifiers?: With Applications to Jury Accuracy, Survey Response Error, and Diagnostic Error

Bruce D. Spencer
Faculty Fellow, Institute for Policy Research
Professor of Statistics
Northwestern University

Version: November 26, 2008; rev. May 7, 2009

Address correspondence to Professor Bruce D. Spencer, Department of Statistics, Northwestern, University, 2006 Sheridan Rd., Evanston IL 60208; email bspencer@northwestern.edu. The author is grateful to Joseph Gastwirth, Shelby Haberman, Wei Li, and Sandy Zabell for helpful comments. Responsibility for errors rests with the author. This research was supported by the Institute for Policy Research, Northwestern University.

Abstract

Latent class models (LCMs) are increasingly used to assess the accuracy of binary classifiers, such as medical diagnostic tests for the presence of a condition, when there is no “gold standard” available and hence the true classification is unknown. LCMs are also used in nonmedical contexts, e.g., for assessing the accuracy of verdicts in criminal cases and for assessing the accuracy of responses to survey questions about drug use, employment status, etc. LCMs posit a relation between observed classifications and unobserved latent classes. When the latent class is treated as the true class, the LCMs provide measures of components of accuracy including the type I and type II error rates. In practice, however, the latent class will differ from the true class and the type I and type II error rates are misspecified by the LCM. A result of Uebersax (1988) implies that under widely applicable conditions, but when covariates are not relevant, LCMs in effect construct the latent class to yield minimum estimates of type I and type II error rates and as a result the LCM estimates of those error rates are optimistic. We derive new lower bounds for the difference between the true type I and type II error rates and those from the LCM; the bounds are applicable and estimable even when covariates are present. The results are important for interpreting the results of latent class analyses. In addition, a total error model is presented that incorporates and error component from invalidity in the LCM.

1. Introduction

Medical diagnostic tests indicate whether a person has a disease, court trials indicate whether a defendant is guilty, surveys report whether a respondent favors of a candidate. These are examples of binary classifiers – procedures that classify individuals with regard to a binary state. There is a substantial and growing literature on how to assess the accuracy of the classifications when the truth is unknown and when a "gold standard" (close approximation treated as truth) is not available. The methods for assessing accuracy of binary classifiers without a gold standard rely on latent class models (*LCMs*), which posit relationships between the observable or *manifest* classification Y and an unobserved or *latent* classification U. When the latent class is used as if it were the true class, V, estimates of accuracy can be obtained. If we denote the classes by 0 or 1, representing respectively absence or presence of the condition of interest (such as a guilt or a disease or preference for candidate X), from the LCM we try estimate the probability of a type I error, $\alpha^{(U)} = P(Y=1|U=0),$ and a type II error, $\beta^{(U)} = P(Y=0|U=1).$ In studies comparing LCM estimates of $\alpha^{(U)}$ and $\beta^{(U)}$ to estimates based on comparing Y to a gold standard, so that $\alpha^{(V)} = P(Y=1|V=0)$ and $\beta^{(V)} = P(Y=0|V=1)$ can be directly estimated, it is frequently found that the estimates of $\alpha^{(U)}$ and $\beta^{(U)}$ are smaller than the corresponding estimates of $\alpha^{(V)}$ and $\beta^{(V)}$. Uebersax (1988, 409) gave conditions and a theoretical explanation why $\alpha^{(U)} \le \alpha^{(V)}$ and $\beta^{(U)} \le \beta^{(V)}$ for homogeneous latent class models, namely models where, for each $u=0,1,$ $P(Y=1|U=u)$ is constant for all cases (Section 2). We extend those results by finding conditions under which $\alpha^{(U)} \le \alpha^{(V)}$ and $\beta^{(U)} \le \beta^{(V)}$ for heterogeneous latent class models, in which $P(Y=1|U=u)$ varies across cases according to random effects and covariates (Section 3). These inequalities are important for understanding the LCM estimates, say $\hat{\alpha}$ and $\hat{\beta},$ of $\alpha^{(V)}$ and $\beta^{(V)}$. We discuss the decomposition of total error in the estimate of type I error, $\hat{\alpha} - \alpha^{(V)},$ into the sum of statistical error $\hat{\alpha} - \alpha^{(U)}$ and invalidity error $\alpha^{(U)} - \alpha^{(V)},$ with a parallel decomposition for type II error (Section 4.) Knowing that the invalidity error is less than or equal to zero is useful in making inferences and decisions based on LCM estimates. All of these results apply directly to specificity, the complement of the type I error probability, and sensitivity, the complement of the type I error probability.

For recent reviews and contributions to the medical literature on evaluating accuracy of diagnostic tests when a gold standard is unavailable, see Hui and Zhou (1998), Pepe and Janes (2007), and Albert and Dodd (2004). Gastwirth and Sinclair (1998) and Spencer (2007) provide estimates of the accuracy of verdicts in criminal cases, where the probability of an erroneous acquittal or erroneous conviction is of interest but the true status of guilt or innocence is unknown. In addition, there has been work on assessing of the accuracy of responses to survey questions without a gold standard; Biemer and Wiesen (2002), Kreuter, Yan, and Tourangeau (2008), Sinclair and Gastwirth (1998), Gastwirth and Sinclair (1996).

2. A Simple Latent Class Model

To describe the simple latent class model, let J denote the number of classifiers and let the observable random variable Y_j denote the classification for the case by classifier j, with $j = 1, \ldots, J$. For the most part, it will be convenient and sufficient to focus on a single case at a time. Error rates on the whole are averages of case-specific error rates, and the inequalities we obtain for the latter will thus apply to the former. Let $\mathbf{Y} = (Y_1, \ldots, Y_J)^T$ denote the vector of binary classifications for the case, where Y_j takes the values 1 or 0 to indicate the presence or absence, respectively, of the condition of interest. Denote the *true* status of the case by V, so that $V = 1$ if the condition truly is present and $V = 0$ if it truly is absent.

The true accuracy of classifier j is defined in terms of $P(Y_j = v \mid V = v)$, and if this probability varies across cases then its average over the population will be used. In many applications it will be more appropriate to think of the accuracy of a classifier type than of a unique classifier, for example in court cases we may consider the accuracy of juries as a group rather than the accuracy of a single jury. The true specificity is $P(Y_j = 0 \mid V = 0)$ and the true probability of a type I error is denoted by $\alpha_j^{(V)} =$ $P(Y_{ij} = 1 \mid V = 0)$. The true sensitivity is $P(Y_j = 1 \mid V = 1)$ and the true probability of type II error is denoted by $\beta_j^{(V)} = P(Y_j = 0 \mid V = 1)$. The marginal distribution of V will be denoted by $\pi_v^V = P(V = v)$. In application, unless a gold standard is present it may not be possible to ascertain the true status V.

LCMs often are used to assess accuracy when the true status is unknown. The law of total probability implies that for any random variable U taking values 0 or 1, the joint probability distribution of $\mathbf{Y}$ may be expressed as $P(\mathbf{Y}) = \sum_{u=0}^{1} P(\mathbf{Y} \mid U = u)\pi_u^U$ with $\pi_u^U = P(U = u).$ In application, U is not directly observed and is referred to as a latent variable, in contrast to the observed or manifest variables $\mathbf{Y}$. We have a *latent class model* if two kinds of conditions hold (Clogg 1995, 317). One condition is that the probability distribution of $\mathbf{Y}$, perhaps conditional on observed covariates, is the same for all cases in the same latent class. If there are no covariates, we may write this homogeneity condition as

(1) $$P(\mathbf{Y} \mid U = u) \text{ is constant for all cases.}$$

The second condition is a form of local independence, a simple example being

(2) $$P(\mathbf{Y} = \mathbf{y}) = \sum_u P(U = u) \prod_{j=1}^{J} P(Y_j = y_j \mid U = u).$$

In many applications, the estimates of accuracy are directed at error rates defined in terms of the latent class U instead of the true class V, i.e., they are directed at $\alpha_j^{(U)} = P(Y_j = 1 \mid U = 0)$ and $\beta_j^{(U)} = P(Y_j = 0 \mid U = 1)$ instead of $\alpha_j^{(V)}$ and $\beta_j^{(V)}$; Uebersax (1988, 409ff), Garrett, Eaton, and Zeger (2002, 1291), Bertrand, Bénichou, Grenier, and Chastang (2005, 697ff). Examples 3 and 4 below compare the error estimates from the LCM with estimates from a gold standard and find that LCM underestimates the error rates. We will say the latent class model is *optimistic* or *optimistic for V* (for classifier j) if

(3) $$\alpha_j^{(V)} \geq \alpha_j^{(U)} \text{ and } \beta_j^{(V)} \geq \beta_j^{(U)}.$$

Remark 1. As an aside, we note that (2) can always be satisfied for binary Y_j's if U is not restricted to be binary valued (Suppes and Zanotti, 1981).

The notation is summarized in the following table.

Table 1. Notation

Y_j	manifest classification of the case by classifier j
U	latent class for the case
V	true or correct class for the case
$\alpha_j^{(U)}$	$P(Y_j = 1 \mid U = 0)$, the probability of a type I error according to the LCM
$\beta_j^{(U)}$	$P(Y_j = 0 \mid U = 1)$, the probability of a type II error according to the LCM
$\alpha_j^{(V)}$	$P(Y_j = 1 \mid V = 0)$, the true probability of a type I error
$\beta_j^{(V)}$	$P(Y_j = 0 \mid V = 1)$, the true probability of a type I error
π_u^U	$P(U = u)$
π_v^V	$P(V = v)$

Example 1. Diagnosis of hearing impairment. Pepe and Janes (2007, 479) present data on 3 tests for hearing impairment applied to 666 individuals, and calculate maximum likelihood estimates of π_0^U, $\alpha_j^{(U)}$, and $\beta_j^{(U)}$, $j = 1, \ldots, 3,$ with state 1 corresponding to disease and 0 to no disease. In addition, a gold standard was present, yielding estimates of $\alpha_j^{(V)}$ and $\beta_j^{(V)}$. The estimates (denoted with ^) of the type I and type II error rates are shown in Table 2.

Table 2. Estimates from Pepe and Janes (2007) for hearing impairment

Test (j)	$\hat{\alpha}_j^{(V)}$ (%)	$\hat{\alpha}_j^{(U)}$ (%)	$\hat{\beta}_j^{(V)}$ (%)	$\hat{\beta}_j^{(U)}$ (%)
1	40.1	12.9	33.6	15.9
2	36.0	13.3	37.5	23.8
3	53.7	31.2	24.9	10.2

Notice that the estimates of type I error and type II based on the LCM are much lower than the estimates according to the gold standard, in accordance with (3). In addition, $\hat{\pi}_1^V = 42\%$ and $\hat{\pi}_1^U = 54\%.$

Example 2. Response errors to alternatively worded survey questionnaire items. Kreuter, Yan, and Tourangeau (2008) compared the results of a latent class analysis with a gold standard analysis in order to identify identifying survey questions that would have large response error. Three alternative survey questions were given in 2005 to individuals who, according to official records had received undergraduate degrees from the University of Maryland between 1989 and 2002: (1) Did you ever receive a grade of 'D' or 'F' for a class? (2) Did you ever receive an unsatisfactory or failing grade? (3) What was the worst grade you ever received in a course as an undergraduate at the University of Maryland? The latent class of interest was whether the student had ever received a grade of D or F as an undergraduate at the University of Maryland, and the true classification was ascertainable from the official records held by the University Registrar. The LCM was (2) with $J = 3$ and with $Y_j = 1$ if the individual indicated having received a failing grade in response to question j. Data were obtained from 954 individuals and statistics for the mean of Y_j and maximum likelihood estimates (denoted with ^) of the type I and type II error rates were calculated under the assumption of conditional independence; see Table 3.

Table 3. Statistics and estimates from Kreuter, Yan, and Tourangeau (2008)

Item (j)	ave. Y_j (%)	$\hat{\alpha}_j^{(V)}$ (%)	$\hat{\alpha}_j^{(U)}$ (%)	$\hat{\beta}_j^{(V)}$ (%)	$\hat{\beta}_j^{(U)}$ (%)
1	45.5	2.9	2.5	26.2	2.1
2	25.4	1.8	1.2	59.0	45.1
3	46.2	2.9	3.2	25.0	1.2

Observe that $\hat{\alpha}_j^{(U)} + \hat{\beta}_j^{(U)} \leq 1$ holds, and (3) holds when the estimated error rates are substituted for the actual probabilities, with the one exception that $\hat{\alpha}_3^{(V)} < \hat{\alpha}_3^{(U)}$. The estimates are close, however, and could reflect random variability. The apparent partial failure of (3) could also arise from model misspecification, including unmodeled covariates, as discussed in the next section.

Proposition 1. (Uebersax 1988) If

$$\alpha_j^{(U)} + \beta_j^{(U)} \leq 1, \tag{4}$$

and

(5) $$P(\mathbf{Y} \mid U,V) = P(\mathbf{Y} \mid U).$$

then (3) holds.

We can expect (4) to hold if the classifiers are not too inaccurate and U is not too poor an approximation to the true latent class, V. To see why, note that a classifier based purely on randomization and no information would have $\alpha_j^{(V)} + \beta_j^{(V)} = \alpha_j^{(U)} + \beta_j^{(U)} = 1$; for a more informative classifier we will have $\alpha_j^{(V)} + \beta_j^{(V)} < 1$ and, if U is not *too* different than V, then we will have $\alpha_j^{(U)} + \beta_j^{(U)} \le 1$ as well. Condition (4) can be checked empirically from data, as in Example 2. Turning to condition (5), note that it is implied by (1). [Proofs are found at the end of the paper.]

Remark 2. Condition (4) is equivalent to the property of *latent monotonicity* (Holland and Rosenbaum 1986, 1525), specifically that $P(Y_j > y \mid U = u)$ is nondecreasing in u for each y, or in the current context of binary variables, $P(Y_j = u \mid U = u) \ge P(Y_j = u \mid U = 1-u)$ for each u. Although the inequalities in (3) do not require that a latent monotonicity property holds between Y_j and V, the inequalities are tighter if the property does hold. Given (4) and (5), a sufficient condition for latent monotonicity between Y_j and V is latent monotonicity between U and V.

Example 3. Hui-Walter model for accuracy of verdicts in criminal trials. Consider a criminal trial by jury to establish the guilt ($V = 1$) or non-guilt ($V = 0$) of a defendant. The true guilt or innocence of the defendant typically is not known, and indeed an important issue for understanding the type I and type II error rates concerns the nature of the "true state" of the defendant. Consider for example a trial of a defendant who committed the crime as charged but for which the evidence is insufficient. Is the correct verdict guilty or not guilty? From an evidentiary or procedural perspective, the correct verdict is not guilty because proof of the crime has not been demonstrated to the standards required by the law. From a material or factual perspective, the correct verdict is is guilty because the defendant actually committed the crime (Laudan 2006, Spencer 2007). A study by Kalven and Zeisel (1966) of more than 3500 jury trials in the 1950s in 47 states and Washington D.C. yielded 411 trials for burglary and auto theft for which a verdict of guilty or not guilty was reported by both the judge ($j=1$) and the jury $(j=2)$. The model (2) is characterized by $1+2J$ parameters, π_0^U, $\alpha_j^{(U)}$, and $\beta_j^{(U)}$, $j=1,\ldots,J$. With only $J=2$ classifiers, the data consist of a 2×2 table with counts of (Y_{i1},Y_{i2}), providing only 3 degrees of freedom to estimate 5 parameters. To be able to estimate the parameters, Gastwirth and Sinclair (1998) adopted a model of Hui-Walter (1980) and considered two types of crime, burglary and auto-theft, for which π_0^U would differ but for which $\alpha_j^{(U)}$ and $\beta_j^{(U)}$ were assumed to be the same. This model employs a pair of LCMs with a restricted parameter space. Gastwirth and Sinclair (1998, 63) assumed the verdicts were conditionally independent within and across cases given the latent class and they developed the following estimates ($s.e.$ denotes estimated standard error): $\alpha_1^{(U)} =$ 0.000 ($s.e. =$ 0.399), $\beta_1^{(U)} =$ 0.012 ($s.e. =$ 0.015), $\alpha_2^{(U)} =$ 0.009 ($s.e. =$ 0.089), $\beta_2^{(U)} =$ 0.192 ($s.e. =$ 0.044), and hence $\alpha_1^{(U)}+\beta_1^{(U)}=$ 0.012 ($s.e. \leq$ 0.414) and $\alpha_2^{(U)}+\beta_2^{(U)} =$ 0.201 ($s.e. \leq$ 0.133). These provide a partial confirmation that (4) holds. One may question whether the model is correctly specified, however, specifically the assumption of constant conditional probabilities of error given U. Cases vary in difficulty and in strength of evidence. Also, agreement between judge and jury has been observed to depend on their demographics (Eisenberg et al 1995), which is consistent with error probabilities varying with the demographic characteristics of classifier and defendant. This likely heterogeneity makes it is plausible that (5) fails to hold, and in that case we would need additional analysis to see whether (3) would hold; Example 8, below, provides further discussion.

Unfortunately, the inequalities in (3) tell us little about the understatement or overstatement of the extents of type I and type II errors, where by *extent* of an error we mean the number of errors as a proportion of all cases. The extent of type I error according to the LCM equals the average across cases of $\pi_0^U\alpha_j^{(U)}$ and the true extent equals the average of $\pi_0^V\alpha_j^{(V)}$; for type II errors the extents are averages across cases of $(1-\pi_0^U)\beta_j^{(U)}$ and $(1-\pi_0^V)\beta_j^{(V)}$, respectively. In Example 1, estimated extents of type I errors and type II errors happened to be smaller according to the LCM than the gold standard; the same ordering applies classifiers 1 and 3 in Example 2 but not for classifier 2.

The extents of type I and type II errors are important for public policy purposes; for example, from a public health standpoint the optimal screening rate for a disease depends on the extents and relative costs of type I and type II errors. In criminal justice there has long been concern with the ratio of the two extents, or the balance of the two kinds of error, specifically that it is worse to convict one innocent person than to acquit n guilty people. The U.S. Supreme Court has interpreted the U.S. Constitution as stipulating that n be much greater than 1; In re Winship, 397 U.S. 358 (1970). Specific prescriptions for n have ranged from 2 (Voltaire) to 1000 (Maimonides), if not more widely (Volokh 1997), although the best-known value may be Blackstone's choice of $n = 10$: "it is better that ten guilty persons escape than that one innocent suffer" (Blackstone 1825, 358). Notice that n refers to the ratio of the extent of type II errors to the extent of type I errors. Even if the inequalities in (3) are strict, we know only that the extent of either type I error or type II error according to the LCM is too low, or that both are, but we do know which. Nor do we know the implications for the ratio of the extents. Thus, we are unable to assess the directional bias for using a LCM instead of a gold standard to empirically estimate the balance of errors.

3. Latent Class Models Allowing Heterogeneity and More General Dependence

3.1. Models without Covariates

More general LCMs have been developed to relax the homogeneity assumption in (1) and the local independence condition in (2). With regard to the latter, Vacek (1985) and Torrance-Rynard and Walter (1997) directly allowed for pair-wise dependence between classifiers, $\text{cov}(Y_j, Y_{j'} \mid U = u) \neq 0$, and Espeland and Handelman (1989) and Yang and Becker (1997) modeled pair-wise dependence via log-linear models. Although such dependence complicates the estimation of accuracy, by itself it does not affect the error probabilities $\alpha_j^{(V)}$, $\beta_j^{(V)}$, $\alpha_j^{(U)}$, $\beta_j^{(U)}$.

The homogeneity assumption in (1) is unrealistic when the conditional probabilities of classification given the latent class U vary across cases. For example, Spencer (2007) found both type I and type II error probabilities to vary with the strength of evidence presented at trial. In a study of exonerations in capital cases in the U.S., Gross and O'Brien (2008) found type I error rates to be positively associated with the length of time from crime to arrest – which is understandable since shorter time lags are associated with easier investigations. To model such heterogeneity, it is convenient to re-express the conditional probability of classification for the case of homogeneous error probabilities as $P(Y_j = 0 | U) = F(-a_j U + b_j)$ with F taken to be any strictly increasing function taking values on the unit interval and having inverse function F^{-1}. This representation entails no loss in generality when the error rates do not vary across cases, as the parameters a_j and b_j are identified as $a_j = F^{-1}(1-\alpha_j^{(U)}) - F^{-1}(\beta_j^{(U)})$ and $b_j = F^{-1}(1-\alpha_j^{(U)})$. (If (4) holds then $a_j \geq 0$.) Given a choice of F, the heterogeneity can be modeled parametrically in terms of a case-specific "effect" R such that $P(Y_j = 0 | R, U = 0) = F(b_j + c_j R)$ and $P(Y_j = 0 | R, U = 1) = F(-a_j + b_j + d_j R)$, where a_j, b_j, c_j, d_j are constant across cases. In random effects models (without covariates), the effects R are random variables distributed independently of the U's according to some mixing distribution, the manifest variables Y_j are conditionally independent given the U's and R's, and some functional form is specified for $P(Y_j = 0 | U, R)$.

Example 4. Random Effects Models. Qu, Tan, and Kutner (1996) proposed and analyzed a Gaussian random effects (GRE) model with

$$P(Y_j = 0 | U, R) = \Phi(-a_j U + b_j + c_j R + (d_j - c_j) U R),$$

where Φ denotes the standard normal distribution function, and the R's have normal distributions and are independent of each other and of the U's. Albert and Dodd (2004) discuss alternative continuous mixing distributions for the R's, as well as discrete mixing distributions as used by Albert *et al.* (2001) and Espeland and Handelman (1989). Hall and Zhou (2003) discuss the theoretical identifiability of such models and Albert and Dodd (2004) discuss difficulties of identifying the models in practice.

Proposition 2 presented in the next subsection implies that when random effects R are present (3) continues to hold if R is independent of V as well as U and conditions (4) and (5) hold conditionally on the random effect.

3.2. Models with covariates

In some applications, the probability distribution of the classification of a case will also depend on a vector of covariates, $\mathbf{X}$.

Example 5. Gaussian Random Effects Model with Covariates. Hadgu and Qu (1998) generalized the GRE model in Example 4 to allow the probabilities to depend on linear functions of covariates $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_K)^T$,

$$P(Y_j = 0 \mid U, R, \mathbf{X}) = \Phi(-a_j U + b_j + c_j R + (d_j - c_j) U R_i + \mathbf{e}_j^T \mathbf{X} + (\mathbf{f}_j^T - \mathbf{e}_j^T) U \mathbf{X})$$

where K is the number of covariates and $\mathbf{e}_j$ and $\mathbf{f}_j$ are $K \times 1$ vectors that do not vary across cases.

It is typically assumed that any random effects R are independent of the U 's and the $\mathbf{X}$'s, and further that the Y_j 's are conditionally independent given the U 's, $\mathbf{X}$'s, and R 's. In some situations, after careful consideration of the nature of the manifest variables, covariates, and concept of true classification, we may assume that the conditional independence properties hold conditional on V as well, specifically that

(6) $$P(R \mid U, V, \mathbf{X}) = P(R),$$

(7) $$P(\mathbf{Y} \mid U, V, R, \mathbf{X}) = P(\mathbf{Y} \mid U, R, \mathbf{X}),$$

(8) $$P(\mathbf{X} \mid U, V) = P(\mathbf{X} \mid U).$$

It often is reasonable to assume that (4) holds conditionally on $\mathbf{X}$,

(9) $$P(Y_j = 1 \mid U = 0, \mathbf{X} = \mathbf{x}) + P(Y_j = 0 \mid U = 1, \mathbf{X} = \mathbf{x}) \leq 1 \quad \text{for all } \mathbf{x}.$$

Proposition 2. If $U, V, R, \mathbf{X}, \mathbf{Y}$ satisfy (6)–(9) then

(10) $$\alpha_j^{(V)} \geq \alpha_j^{(U)} + \min\{0, A_j\}, \qquad \beta_j^{(V)} \geq \beta_j^{(U)} + \min\{0, B_j\}$$

with $A_j = \sum_{\mathbf{x}} P(Y_j = 1 \mid U = 0, \mathbf{X} = \mathbf{x}) \Delta(U, X)$, $B_j = -\sum_{\mathbf{x}} P(Y_j = 0 \mid U = 1, \mathbf{X} = \mathbf{x}) \Delta(U, X)$, and

$\Delta(U,X) = P(\mathbf{X} = \mathbf{x} \mid U = 1) - P(\mathbf{X} = \mathbf{x} \mid U = 0).$

The proposition implies that (3) holds and the model is optimistic with respect to V if $A_j \geq 0$, $B_j \geq 0$. Note that A_j and B_j can be estimated from the data, as in Example 7 below, and the precision of those estimates can also be estimated from the data.

Example 6. Importance of Condition (8). It is possible for (3) to fail badly when condition (8) does not hold. Suppose a covariate X takes values 0 and 1, and $P(Y_j = X \mid U,V) = P(Y_j = X) = 1;$ that is, the covariate completely determines the classification. Suppose the joint distribution over (X,U,V) is

$$P(X = x, U = u, V = v) = \begin{cases} r & \text{if } X = 1, U = 0, V = 1 \\ 1 - r - \varepsilon & \text{if } X = 1, U = 1, V = 1 \\ \varepsilon/6 & \text{otherwise,} \end{cases}$$

with ε a small positive number and $r \geq 0.5$. Condition (8) fails to hold, since $P(X = 0 \mid U = 1, V = 0) \neq P(X = 0 \mid U = 1, V = 1),$ but the other conditions of the proposition do apply and in particular $A_j \geq 0$, $B_j \geq 0$. We have $\alpha_j^{(U)} \approx 1$ but $\alpha_j^{(V)} \approx 0.5;$ here the latent class model would grossly *overstate* the type I error rate. We also have $\beta_j^{(U)} \approx 0$ but $\beta_j^{(V)} \approx$ $(1-r)\beta_j^{(U)},$ so the latent class model would also (very slightly) overstate the type II error rate.

Example 7. Log-linear Latent Class Model with Covariates. Spencer (2007) analyzed the accuracy of verdicts in a convenience sample of criminal trials in 2000-01 in four jurisdictions in the U.S., Los Angeles, Washington D.C., the Bronx, and Maricopa County, Arizona (which includes the city of Phoenix). Data were available for 271 trials showing the verdict that the judge reported he or she would have issued had the trial been a bench trial (Y_1), the jury's verdict (Y_2), a 3 point rating by the judge of the strength of evidence for conviction (X_1), and an analogous 3 point rating by the jury (X_2). The data observed data come from a table with 36 cells such that each cell includes both values of the unobserved latent variable U. The observed 36 cell table was treated as a partially observed 72 cell table derived by splitting each of the 36 observed cells U. The data were treated as arising from a multinomial distribution where the log of the probability of falling in any of the 72 cells was specified by an ANOVA model involving main effects for $U,\ X_1,\ X_2,\ X_1, Y_2$ and pair-wise interactions between U and $X_1,\ X_2,\ Y_1,\ Y_2$, between X_1 and X_1, and between X_2 and Y_2. Maximum likelihood estimates of the error probabilities were $\hat{\alpha}_1^{(U)} = 0.37$ and $\hat{\beta}_1^{(U)} = 0.02$ for the judge and $\hat{\alpha}_2^{(U)} = 0.25$ and $\hat{\beta}_2^{(U)} = 0.14$ for the jury. The type I error rates are surprisingly large, and one may wonder how they would change if the latent class represented *true* guilt or innocence V as described in Example 3. The maximum likelihood estimates of A_j and B_j are, respectively, $\hat{A}_1 = -0.18$ and $\hat{B}_1 = 0.06$ for the judge $(j = 1)$ and $\hat{A}_2 = -0.003$ and $\hat{B}_2 = 0.12$ for the jury $(j = 2)$. Although conditions (7) and (8) are not directly testable, it is plausible that they hold because $\mathbf{X}$ is determined by the evaluations by judges and juries of the strength of evidence, and $\mathbf{Y}$ represents judgements based on evaluations the evidence, U is empirically determined by $\mathbf{X}$ and $\mathbf{Y}$ in the context of the model; it is not apparent that the correct verdict V would affect either the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ and U or the conditional distribution of $\mathbf{X}$ given U. Assuming, then, that the conditions of Proposition 2 hold, we infer that $\alpha_2^{(V)}$ is as large as $\alpha_2^{(U)}$ or not too much smaller (allowing for estimation error in $\hat{A}_2$). Practically, we may take the jury's estimated type I error rate from the latent class model as an optimistic estimate of the error rate defined with respect to the *correct* verdict V, however defined. For the judge, with $\hat{A}_1 = -0.18$, we see that because V and U may differ the large estimated type I error rate $(\hat{\alpha}_1^{(U)} = 0.37)$ *could* be an overstatement of $\alpha_2^{(U)}$ even if $\hat{\alpha}_1^{(U)} \le \alpha_1^{(U)}$, but the possibility that it is an understatement cannot be excluded either. Turning to type II errors, the fact that $\hat{B}_j > 0$ suggests that $\beta_j^{(V)} \ge \beta_j^{(U)}$ for $j = 1, 2$. Spencer (2007, 315, 323) cautions against generalizing from the estimates due to the nonrandom nature of the sample and lack of sampling weights, among other considerations.

4. Total Error Decomposition

Suppose that $U, V, R, \mathbf{X}, \mathbf{Y}$ satisfy (6)–(9), and suppose V is the true state. We allow for the possibility that the latent class model used to estimate $\alpha^{(V)}$ and $\beta^{(V)}$ be misspecified. The misspecification might arise from omission of covariates, the wrong distribution for random effects, failure to account for measurement error, etc. The estimation model leads to estimates $\hat{\alpha}$ and $\hat{\beta}$. Suppose that the estimates have limiting values as sample sizes increase, say $\tilde{\alpha}$ and

$\tilde{\beta}$. The difference between the estimate of the type I error rate and the true value of the error rate, $\hat{\alpha} - \alpha^{(V)}$, is the total error in the estimate. The total error is equal to the sum of three components, $\hat{\alpha} - \tilde{\alpha}$, $\tilde{\alpha} - \alpha^{(U)}$, and $\alpha^{(U)} - \alpha^{(V)}$. The first component represents sampling error and the second error represents non-sampling error, including misspecification of the estimation model and problems in the sampling frame used for selecting cases. The third component depends in part on the intended use of the classification and correspondingly on the intended use the estimate of accuracy. In consonance with the psychometric literature, we may call the third component, $\alpha^{(U)} - \alpha^{(V)}$, invalidity. For example, in discussing the validity of test scores, Linn (1983, 336ff) noted that "Validity is not a property of a test that is independent of the uses that are made of it. It is interpretation of the test results for a given population that is validated rather than the test in isolation. Thus, a single test may have many validities associated with the many interpretations made of its results . . ." A parallel decomposition applies to type II error rates, giving us the following decompositions.

$$\hat{\alpha} - \alpha^{(V)} = \underbrace{\hat{\alpha} - \tilde{\alpha}}_{\substack{\text{sampling}\\ \text{error}}} + \underbrace{\tilde{\alpha} - \alpha^{(U)}}_{\substack{\text{non-sampling}\\ \text{error}}} + \underbrace{\alpha^{(U)} - \alpha^{(V)}}_{\text{invalidity}},$$

$$\hat{\beta} - \beta^{(V)} = \underbrace{\hat{\beta} - \tilde{\beta}}_{\substack{\text{sampling}\\ \text{error}}} + \underbrace{\tilde{\beta} - \beta^{(U)}}_{\substack{\text{non-sampling}\\ \text{error}}} + \underbrace{\beta^{(U)} - \beta^{(V)}}_{\text{invalidity}}.$$

Remark 3. Although validity often is measured with a correlation coefficient, the justification for that practice is narrow and is justified from the use of measurements for selecting individuals (for a job, admission to school, promotion, etc.), as in Cochran (1951) and Cronbach and Gleser (1957). To illustrate, let $\mathbf{X}$ denote the measurements on the individual, let Z denote the utility derived if the individual is selected, and define $t(\mathbf{X}) = E[Z \mid \mathbf{X}]$. Such a Z is called the *criterion*. The value from selecting Individuals with $t(\mathbf{X})$ exceeding a threshold are selected. Assuming the distribution of $t(\mathbf{X})$ belongs to a scale family, we can show that the value of selecting in that way based on $t(\mathbf{X})$ compared to selecting individuals at random is then proportional to the correlation between $t(\mathbf{X})$ and Z, say ρ_{Zt}, and this correlation is called the *criterion validity*. Let θ denote a latent trait of the individual that is sufficient for the selection problem, in the sense that Z and $E[Z \mid \mathbf{X}]$ are uncorrelated given θ. Define $\tau = E[t(\mathbf{X}) \mid \theta]$; τ is called a *construct*. The *construct validity* of the measure t is defined as the correlation between τ and $(t\ \mathbf{X}),$ say

$\rho_{\tau t}$. The *criterion validity of the construct* is defined as the correlation between Z and τ, say $\rho_{Z\tau}$. Thus, the criterion validity factors into the product of the criterion validity of the construct and the construct validity, $\rho_{Zt} = \rho_{Z\tau}\,\rho_{\tau t}$. It should be appreciated that the construct is defined in the context of the selection problem, that is, with respect to utility, and that whole development of correlations as measures of validity is particular to the selection problem and need not be more widely appropriate.

Example 8. Hui-Walter model for accuracy of verdicts in criminal trials (continued). In the discussion of Example 3 it was suggested that the Hui-Walter model as employed by Gastwirth and Sinclair (1998) was likely misspecified. However, it is plausible that, for some covariate $\mathbf{X}$, conditions (6)–(9) hold for $U, V, R, \mathbf{X}, \mathbf{Y}$. Although the model misspecification will affect $\tilde{\alpha} - \alpha^{(U)}$ and $\tilde{\beta} - \beta^{(U)}$ in an as yet unknown way, Proposition 2 implies that the validity error can be bounded below by (11). To estimate (12) retrospectively from the available information is quite difficult. As an illustration, we can try using the results from the more recent study as discussed in Example 7. This would lead us to conclude that for all practical purposes $\beta_1^{(V)} \geq \beta_1^{(U)}$, $\beta_2^{(V)} \geq \beta_2^{(U)}$, and $\alpha_2^{(V)} \geq \alpha_2^{(U)}$, but that we cannot say much about the difference between $\alpha_1^{(V)}$ and $\alpha_1^{(U)}$. The standard error on the estimate of α_1, at 0.4, is so extremely large that the sampling error likely dominates the invalidity error for this parameter.

5. Conclusions

When estimates of type I and type II error rates for binary classifiers are based on latent class models (LCMs), the estimates themselves have a component of error attributable to the fact that the latent class and the true class of interest are not the same. The difference is in effect a type of invalidity, and under certain conditions it can be quantified via lower bounds on the difference between the type I or type II error rate for the true class and the corresponding error rate for the latent class operationally defined by the LCM. The lower bounds are estimable from the data used to fit the LCM. In many cases, the lower bound is zero, which means that the invalidity in the latent class contributes a non-negative component of bias to the estimates of type I and type II error rates. These results help explain the empirical findings that when LCM estimates of type I and type II error rates are compared to those based on comparisons to gold standards, the LCM error rates tend to be too small.

References

Albert, P. S. and Dodd, L. E. (2004) A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, **60**, 427-435.

Albert, P. S., McShane, L. M., Shih, J. H. and the U.S. National Cancer Institute Bladder Tumor Marker Network (2001) Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. *Biometrics,* **57**, 610-619.

Bertrand, P., Bénichou, J., Grenier, P. and Chastang, C. (2005). Hui and Walter's latent-class reference-free approach may be more useful in assessing agreement than diagnostic performance. *Journal of Clinical Epidemiology*, **58**, 688-700.

Biemer, P. P. and Wiesen, C. (2002) Measurement error evaluation of self-reported drug use: a latent class Analysis of the US National Household Survey on Drug Abuse. *Journal of the Royal Statistical Society. Series A*, **165**, 97-119.

Blackstone, W. (1825) *Commentaries on the Laws of England. Book the Fourth. 16 ed.* London: Strahan.

Clogg, C. C. (1995) Latent class models. Pp. 311-359 in G. Arminger, C.C. Clogg, and M.E. Sobel (eds.) *Handbook of Statistical Modeling for the Social and Behavioral Sciences.* New York: Plenum.

Cochran, W. G. (1951) Improvement by means of selection. Pp. 449-470 in J. Neyman (Ed.), *Second Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press.

Cronbach, L. J. and Gleser, G. C. (1957) *Psychological Tests and Personnel Decisions*. Urbana: University of Illinois Press.

Eisenberg T., Hannaford-Agor P. L., Hans V. P., Waters N. L., Munsterman G. T., Schwab S. J., and Wells M. T. (2005) Judge-jury agreement in criminal cases: a partial replication of Kalven and Zeisel's *The American Jury*. *Journal of Empirical Legal Studies*, **2**, 171-206.

Espeland, M. A. and Handelman, S. L. (1989) Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics,* **45**, 587-599.

Garrett, E. S., Eaton, W. W. and Zeger, S. (2002) Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: a latent class model approach. *Statistics in Medicine,* **21**, 1289–1307.

Gastwirth, J. L. and Sinclair, M. D. (1996) On procedures for evaluating the effectiveness of reinterview survey methods: application to labor force data. *Journal of the American Statistical Association,* **91**, 961-969.

Gastwirth, J. L. and Sinclair, M. D. (1998) Diagnostic test methodology in the design and analysis of judge-jury agreement studies. *Jurimetrics*, **39**, 59-78.

Gastwirth, J. L. and Sinclair, M. D. (2004) A re-examination of the 1966 Kalven-Zeisel study of judge-jury agreements and disagreements and their Causes. *Law, Probability and Risk,* **3**, 169-191.

Gross, S. R. and O'Brien, B. (2008) Frequency and predictors of false conviction: why we know so little, and new data on capital cases. *Journal of Empirical Legal Studies,* **4**, 927–962.

Haberman S. J. (1979) *Analysis of Qualitative Data, Vol. 2.* New York: Academic Press.

Hadgu, A. and Qu, Y. (1998). A biomedical application of latent class models with random effects. *Applied Statistics,* **47**, 606–616.

Hall, P. and Zhou, X.-H. (2003) Nonparametric estimation of component distributions in a multivariate mixture. *The Annals of Statistics,* **31**, 201–224.

Holland, P. W. and Rosenbaum, P. R. (1986) Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, **14**, 1523-1543.

Hui, S. L. and Walter S. D. (1980) Estimating the error rates of diagnostic tests. *Biometrics,* **36**, 167-171.

Hui, S. L. and Zhou, X. H. (1998) Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research,* **7**, 354-370.

Kalven, H., and Zeisel, H. (1966) *The American Jury*. Boston: Little, Brown.

Kreuter, F., Yan, T. and Tourangeau, R. (2008) Good item or bad—can latent class analysis tell?: the utility of latent class analysis for the evaluation of survey questions. *Journal of the Royal Statistical Society. Series A*, **171**, 723-738.

Laudan, L. (2006) *Truth, Error, and Criminal Law: An Essay in Legal Epistemology*. New York: Cambridge University Press.

Pepe, M. S. and Janes, H. (2007) Insights into latent class analysis of diagnostic test performance. *Biostatistics*, **8**, 474–484.

Qu, Y., Tan, M. and Kutner, M. H. (1996) Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, **52**, 797-810.

Simpson, E. H. (1951) The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B*, **13**, 238-241.

Sinclair, M. D. and Gastwirth, J. L. (1998) Estimates of the errors in classification in the labour force survey and their effect on the reported unemployment rate. *Survey Methodology*, **24**, 157-169.

Spencer, B. D. (2007) Estimating the accuracy of jury verdicts. *Journal of Empirical Legal Studies*, **4**, 305–329.

Suppes, P. and Zanotti, M. (1981) When are probabilistic explanations possible? *Synthese*, **48**, 191-199.

Torrance-Rynard, V. L. and Walter, S. D. (1997) Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine*, **16**, 2157-2175.

Uebersax, J. S. (1988) Validity inferences from interobserver agreement. *Psychological Bulletin*, **104**, 405-416.

Vacek, P. M. (1985) The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, **41**, 959-968.

Volokh, A. (1977) n guilty men. *University of Pennsylvania Law Review*, **146**, 173-216.

Walter S. D. and Irwig, L. M. (1988) Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Clinical Epidemiology,* **41**, 923-937.

Yang, I. and Becker, M.P. (1997) Latent variable modeling of diagnostic accuracy. *Biometrics,* **53**, 948-958.

Detailed Proofs

Proof of Proposition 1.

Uebersax (1988) noted that $\alpha_j^{(V)} = P(Y_j = 1 | V = 0)$ is a weighted average of $P(Y_j = 1 | V = 0, U = 0)$ and $P(Y_j = 1 | V = 0, U = 1)$, and by (4) that equals the same weighted average of $P(Y_j = 1 | U = 0) = \alpha_j^{(U)}$ and $P(Y_j = 1 | U = 1) = 1 - \beta_j^{(U)}$. But $1 - \beta_j^{(U)} \ \geq \alpha_j^{(U)}$ by (4), and hence $\alpha_j^{(V)} \geq \alpha_j^{(U)}$. Similarly, $\beta_j^{(V)} \geq \beta_j^{(U)}$. Thus, (3) is established.

Proof of Remark 2.

Remark 2 makes 3 assertions. To prove the first, that (4) is equivalent to latent monotonicity, suppose first that latent monotonicity holds, and notice that $1 - \beta_j^{(U)} = P(Y_j = 1 | U = 1) \geq$ $P(Y_j = 1 | U = 0) = \alpha_j^{(U)}$, establishing (4). On the other hand, if (4) holds then

$$0 \leq 1 - \alpha_j^{(U)} - \beta_j^{(U)} = 1 - P(Y_j = 1 | U = 0) - P(Y_j = 0 | U = 1) = P(Y_j = 0 | U = 0) - P(Y_j = 0 | U = 1)$$

$$\Rightarrow P(Y_j = 0 | U = 1) \leq P(Y_j = 0 | U = 0)$$

and

$$0 \leq 1 - \alpha_j^{(U)} - \beta_j^{(U)} = 1 - P(Y_j = 0 | U = 1) - P(Y_j = 1 | U = 0) = P(Y_j = 1 | U = 1) - P(Y_j = 1 | U = 0)$$

$$\Rightarrow P(Y_j = 1 | U = 0) \leq P(Y_j = 1 | U = 1)$$

so $P(Y_j = u | U = u) \geq P(Y_j = u | U = 1 - u)$ for each $u,$ and latent monotonicity is established.

Next, to see that the inequalities in (3) are tighter if the latent monotonicity property holds between Y_j and $V,$ suppose that (4), (5), and hence (3) hold but latent monotonicity does not hold between Y_j and V. Then latent monotonicity holds between Y_j and $\tilde{V} = 1 - V$, and we have $\beta_j^{(V)} = 1 - P(Y_j = 1 | \tilde{V} = 0) \geq 1 - P(Y_j = 1 | \tilde{V} = 1) = \beta_j^{(\tilde{V})} \geq \beta_j^{(U)}$; similarly $\alpha_j^{(V)} \geq \alpha_j^{(\tilde{V})} \geq \alpha_j^{(U)}$.

Finally, we show if (4) and (5) hold, latent monotonicity between U and V implies latent monotonicity between Y_j and V. First observe that for the current context with binary variables there is latent monotonicity between Y_j and V if and only if $P(Y_j = u | V = u) - P(Y_j = u) \geq 0$. Notice that

$$P(Y_j = u | V = u) = P(Y_j = u | U = u, V = u) P(U = u | V = u)$$

$$+ P(Y_j = u | U = 1 - u, V = u) P(U = 1 - u | V = u)$$

$$= P(Y_j = u | U = u) P(U = u | V = u) + P(Y_j = u | U = 1 - u) P(U = 1 - u | V = u)$$

$$= P(Y_j = u | U = u) P(U = u | V = u) + P(Y_j = u | U = 1 - u)[1 - P(U = u | V = u)]$$

and

$$P(Y_j = u) = P(Y_j = u | U = u) P(U = u) + P(Y_j = u | U = 1 - u) P(U = 1 - u)$$

$$= P(Y_j = u | U = u) P(U = u) + P(Y_j = u | U = 1 - u)[1 - P(U = u)].$$

Subtracting, we have

$$\begin{aligned}
P(Y_j = u \mid V = u) - P(Y_j = u) &= P(Y_j = u \mid U = u)P(U = u \mid V = u) \\
&\quad + P(Y_j = u \mid U = 1-u)[1 - P(U = u \mid V = u)] \\
&\quad - P(Y_j = u \mid U = u)P(U = u) + P(Y = u \mid U = 1-u)[1 - P(U = u)] \\
&= P(Y_j = u \mid U = u)[P(U = u \mid V = u) - P(U = u)] \\
&\quad + P(Y_j = u \mid U = 1-u)[P(U = u) - P(U = u \mid V = u)] \\
&= [P(Y_j = u \mid U = u) - P(Y_j = u \mid U = 1-u)][P(U = u \mid V = u) - P(U = u)] \\
&\geq 0.
\end{aligned}$$

Thus, there is latent monotonicity between Y_j and V.

Detailed Proof of Proposition 1

We use a simple conditioning argument, as in Uebersax (1988) and Bertrand, Bénichou, Grenier, and Chastang (2005). We will use the notation E_R to denote expectation with respect to the distribution of $R,$ $E_{\mathbf{X}|V=0}$ to denote expectation with respect to the conditional distribution of $\mathbf{X}$ given $V = 0,$ etc. Observe first that

$$\begin{aligned}
\alpha_j^{(V)} &= P(Y_j = 1 \,|\, V = 0) \\
&= E_{\mathbf{X}|V=0} E_{R|\mathbf{X},V=0} P(Y_j = 1 \,|\, V = 0, \mathbf{X}, R) \\
&= E_{\mathbf{X}|V=0} E_{R|\mathbf{X},V=0} \sum_{u=0}^{1} P(Y_j = 1 \,|\, U = u, V = 0, \mathbf{X}, R) P(U = u \,|\, V = 0, \mathbf{X}, R) \\
&\overset{\text{by (7)}}{=} E_{\mathbf{X}|V=0} E_R \sum_{u=0}^{1} P(Y_j = 1 \,|\, U = u, V = 0, \mathbf{X}, R) P(U = u \,|\, V = 0, \mathbf{X}) \\
&\overset{\text{by (8)}}{=} E_{\mathbf{X}|V=0} E_R \sum_{u=0}^{1} P(Y_j = 1 \,|\, U = u, \mathbf{X}, R) P(U = u \,|\, V = 0, \mathbf{X}) \\
&\overset{\text{by (7)}}{=} E_{\mathbf{X}|V=0} \sum_{u=0}^{1} P(Y_j = 1 \,|\, U = u, \mathbf{X}) P(U = u \,|\, V = 0, \mathbf{X}) \\
&= E_{\mathbf{X}|V=0} \left\{ P(Y_j = 1 \,|\, U = 0, \mathbf{X}) P(U = 0 \,|\, V = 0, \mathbf{X}) + P(Y_j = 1 \,|\, U = 1, \mathbf{X}) P(U = 1 \,|\, V = 0, \mathbf{X}) \right\} \\
&\overset{\text{by (12)}}{\geq} E_{\mathbf{X}|V=0} \left\{ P(Y_j = 1 \,|\, U = 0, \mathbf{X}) P(U = 0 \,|\, V = 0, \mathbf{X}) + P(Y_j = 1 \,|\, U = 0, \mathbf{X}) P(U = 1 \,|\, V = 0, \mathbf{X}) \right\} \\
&= E_{\mathbf{X}|V=0} P(Y_j = 1 \,|\, U = 0, \mathbf{X}).
\end{aligned}$$

In contrast, $\alpha_j^{(U)} = E_{\mathbf{X}|U=0} P(Y_j = 1 \,|\, U = 0, \mathbf{X})$ and so we have

$$\begin{aligned}
\alpha_j^{(V)} - \alpha_j^{(U)} &\geq E_{\mathbf{X}|V=0} P(Y_j = 1 \,|\, U = 0, \mathbf{X}) - E_{\mathbf{X}|U=0} P(Y_j = 1 \,|\, U = 0, \mathbf{X}) \\
&= \sum_{\mathbf{x}} P(Y_j = 1 \,|\, U = 0, \mathbf{X} = \mathbf{x}) \left[P(\mathbf{X} = \mathbf{x} \,|\, V = 0) - P(\mathbf{X} = \mathbf{x} \,|\, U = 0) \right]
\end{aligned}$$

where for simplicity of exposition we treat the covariates as discrete. Notice that

$$\begin{aligned}
P(\mathbf{X} = \mathbf{x} \,|\, V = 0) &= P(\mathbf{X} = \mathbf{x} \,|\, U = 0, V = 0) P(U = 0 \,|\, V = 0) + P(\mathbf{X} = \mathbf{x} \,|\, U = 1, V = 0) P(U = 1 \,|\, V = 0) \\
&\overset{\text{by (9)}}{=} P(\mathbf{X} = \mathbf{x} \,|\, U = 0) P(U = 0 \,|\, V = 0) + P(\mathbf{X} = \mathbf{x} \,|\, U = 1) P(U = 1 \,|\, V = 0)
\end{aligned}$$

and so

$$P(\mathbf{X}=\mathbf{x}\,|\,V=0)-P(\mathbf{X}=\mathbf{x}\,|\,U=0)=P(\mathbf{X}=\mathbf{x}\,|\,U=0)[P(U=0\,|\,V=0)-1]+P(\mathbf{X}=\mathbf{x}\,|\,U=1)P(U=1\,|\,V=0)$$

$$=P(\mathbf{X}=\mathbf{x}\,|\,U=0)[-P(U=1\,|\,V=0)]+P(\mathbf{X}=\mathbf{x}\,|\,U=1)P(U=1\,|\,V=0)$$

$$=[P(U=1\,|\,V=0)][P(\mathbf{X}=\mathbf{x}\,|\,U=1)-P(\mathbf{X}=\mathbf{x}\,|\,U=0)].$$

It follows that

$$\alpha_j^{(V)}-\alpha_j^{(U)}\geq\sum_{\mathbf{x}}P(Y_j=1\,|\,U=0,\mathbf{X}=\mathbf{x})\big[P(\mathbf{X}=\mathbf{x}\,|\,V=0)-P(\mathbf{X}=\mathbf{x}\,|\,U=0)\big]$$

$$=[P(U=1\,|\,V=0)]\sum_{\mathbf{x}}P(Y_j=1\,|\,U=0,\mathbf{X}=\mathbf{x})[P(\mathbf{X}=\mathbf{x}\,|\,U=1)-P(\mathbf{X}=\mathbf{x}\,|\,U=0)].$$

Thus, a sufficient condition for (6) is that

$$0\leq\sum_{\mathbf{x}}P(Y_j=1\,|\,U=0,\mathbf{X}=\mathbf{x})[P(\mathbf{X}=\mathbf{x}\,|\,U=1)-P(\mathbf{X}=\mathbf{x}\,|\,U=0)].$$

This condition is empirically testable under the model.

The condition for $\beta_j^{(V)}\geq\beta_j^{(U)}$ follows similarly.