UW Biostatistics Working Paper Series

1-7-2005

Insights into Latent Class Analysis

Margaret S. Pepe
University of Washington, mspepe@u.washington.edu

Holly Janes
University of Washington, hjanes@u.washington.edu

1. INTRODUCTION

Assessments of the presence or absence of a condition cannot always be made with certainty. This is particularly true in the development of new diagnostic tests, where the very reason a new test is being developed is often because the best available test for the condition is not considered adequately accurate. A problem then arises: How can the accuracy of a new test be evaluated when there is no gold standard against which to compare it? Latent class analysis has been proposed as a statistical technique that allows such an assessment (Walter and Irwig, 1988; Dawid and Skene, 1979). Briefly, a probabilistic model is assumed for the relationship between the new diagnostic test, one or more imperfect "reference" tests, and the unobserved, or latent, disease status. The likelihood is then maximized to provide estimates of the sensitivity and specificity of the new diagnostic test. This approach is quite popular. Recently, it has been used to study markers of Behcet's disease (Ferraz et al., 1995), gastro-oesophageal reflux disease (Moayyedi et al., 2004), visceral leishmaniasis (Boelaert et al., 2004), and acute bacterial rhinosinusitis (Young et al., 2003). Moreover it has received substantial attention from statistical methodologists to extend its applications (Yang and Becker, 1997; Qu et al., 1996; Dendukuri and Joseph, 2001; Hui and Zhou, 1998).

The latent class approach has been criticized on several grounds (Pepe and Alonzo, 2001; Pepe, 2003 [pp 203–205]; Albert and Dodd, 2004). First, the approach yields estimates of the accuracy with which the test predicts disease status, despite the fact that disease is not clinically defined. This means that the estimates of test accuracy themselves are not well-defined. Second, the assumed latent class model is not fully testable with the observed data, and, if the model is incorrect, it is not clear that the resulting estimates are meaningful. Third, the latent class estimates of test accuracy are obtained through a sort of 'black-box' procedure; it is not clear what these estimates are in terms of the raw data. In this paper, we address this third criticism by deriving analytic forms for the estimators. This is particularly useful for assessing their merit when the assumed latent class structure fails, and leads to some general implications for the role of latent class analysis in practice.

The paper is organized as follows. We first describe the latent class analysis technique, and the conditional independence assumption on which the classical latent class model is based. In Section 3, we provide analytical forms for the latent class estimators, and discuss their validity both when the conditional independence assumption holds, and when it fails to hold. The EM-algorithm is used in Section 4 to demonstrate the relationships among the parameter estimates, and to provide expressions which allow us to assess the bias in the estimates caused by conditional dependence. In Sections 3 and 4, we focus on the special case where three tests are available. We conclude in Section 5 that our results lead us to caution against the use of latent class analysis in general.

2. LATENT CLASS ANALYSIS

Classic latent class analysis (LCA) is briefly described as follows: let the binary variable D indicate the presence (D=1) or absence (D=0) of the condition. This is the unobservable latent class variable. The data we observe are the results of K binary test variables, $\{Y_1, \ldots, Y_K\}$, for each of $i = 1, \ldots, n$ subjects. One of these variables may be the best available reference test, and others may be new tests. A statistical model with parameters θ is assumed for the joint distribution of $\{Y_1, \ldots, Y_K\}$ given D, denoted by $P_\theta(Y_1, \ldots, Y_K|D)$. If the model has sufficient structure, θ and the prevalence, $\rho = P(D = 1)$, can be estimated by maximizing the likelihood function

$$\mathcal{L}(\theta, \rho) = \prod_{i=1}^{n} \{\rho P_\theta(Y_{i1}, \ldots Y_{iK}|D = 1) + (1 - \rho) P_\theta(Y_{i1}, \ldots Y_{iK}|D = 0)\}\,.$$

The simplest and most popular statistical model for $P_\theta(Y_1, \ldots, Y_K|D)$ assumes that, given true status, D, the test variables $\{Y_1, \ldots, Y_K\}$ are statistically independent. This is called the *conditional independence* (CI) assumption. It yields the likelihood

$$\mathcal{L}(\theta, \rho) = \prod_{i=1}^{n} \left\{ \rho \prod_{k=1}^{K} P(Y_{ik}|D=1) + (1-\rho) \prod_{k=1}^{K} P(Y_{ik}|D=0) \right\} ,$$

where the parameters, θ, are the true- and false-positive rates, $\phi_k = P(Y_{ik} = 1|D = 1)$, $\psi_k = P(Y_{ik} = 1|D = 0)$, and $\theta = \{(\phi_k, \psi_k),\ k = 1, \dots K\}$. The parameters ϕ_k and ψ_k are also known, respectively, as the sensitivity and (1−specificity) of the kth test. It turns out that, with a minimum of $K = 3$ observed tests, the CI likelihood can be maximized with respect to $\theta = \{(\phi_k, \psi_k), k = 1, \dots, K\}$ and ρ.

The CI assumption is the keystone of the classical latent class approach. The assumption states that, conditional on disease status, the results of the K tests are independent, and knowledge of one test result gives no information about other test results. It is widely acknowledged that this assumption is likely to fail in many cases. For example, if the tests are designed to detect a particular substance in a biological sample, the amount of the substance present in the sample will affect all test results. In many other cases, diagnostic tests are correlated due to disease severity; highly diseased subjects who test positive with one test are likely to test positive with the others. The tests may be independent among controls, but not among cases. We will focus our discussion in the next sections on the setting where $K = 3$ tests are available. In that setting, it is important to note that the validity of the CI assumption cannot be determined at all from the data. When $K \geq 4$, the dependence structure can be modeled. A wide variety of approaches have been taken (Yang and Becker, 1997; Qu et al., 1996; Albert et al., 2001; Espeland and Handelman, 1989). However, Albert and Dodd (2004) have shown that, typically, it is impossible to discern one form of dependence structure from the other.

To illustrate classical LCA, consider the data shown in Table 1 for three tests of hearing impairment measured on $n = 666$ subjects, reproduced from Pepe, 2003 (page 201). The maximum likelihood estimates of the 7 parameters, ρ =prevalence and (ϕ_k, ψ_k) for each of the three tests, are also shown.

Table 1 here

3. ANALYTIC EXPRESSIONS FOR ESTIMATES

Pepe and Alonzo (2001) criticized LCA on the grounds that the connection between the observed data and the parameter estimates is not explicit. An intuition for the estimates does not exist. Practitioners who are not knowledgeable about likelihood functions might simply have faith in the validity of this statistical methodology. Indeed, even for those of us who understand likelihood based methods, the lack of explicit expressions for the parameter estimates in terms of the raw data makes connections to the raw data elusive. Here we rectify this state of affairs, for the special case where $K = 3$ tests are available, by deriving analytic expressions for the estimates in terms of the raw data. Implications of our results for $K \geq 4$ will be discussed in Section 5.

Suppose that there are $K = 3$ observed tests, and write the probabilities of observable data with the following notation:

$$p_k = P(Y_k = 1), \quad k = 1, 2, 3$$

$$p_{kj} = P(Y_k = 1, Y_j = 1), \quad j > k$$

$$p_{123} = P(Y_1 = 1, Y_2 = 1, Y_3 = 1) \; .$$

The same notation with a 'hat' denotes the observed frequency, e.g., $\widehat{p}_k$ is the proportion of observations with $Y_k = 1$. In the appendix we derive the following analytic expressions for the LCA parameter estimates

$$\widehat{\phi}_k = \widehat{p}_k + \sqrt{\widehat{C}_k}\sqrt{\frac{1-\widehat{\rho}}{\widehat{\rho}}} \qquad (1)$$

$$\widehat{\psi}_k = \widehat{p}_k - \sqrt{\widehat{C}_k}\sqrt{\frac{\widehat{\rho}}{1-\widehat{\rho}}} \qquad (2)$$

where

$$\begin{aligned} C_k = & \left(\frac{p_{kj}-p_kp_j}{p_kp_j}\right)\left(\frac{p_{kl}-p_kp_l}{p_kp_l}\right)/\left(\frac{p_{jl}-p_jp_l}{p_jp_l}\right) \\ =& E\left[\left(\frac{Y_k-p_k}{p_k}\right)\left(\frac{Y_j-p_j}{p_j}\right)\right]E\left[(\frac{Y_k-p_k}{p_k})(\frac{Y_l-p_l}{p_l})\right]/E\left[(\frac{Y_j-p_j}{p_j})(\frac{Y_l-p_l}{p_l})\right] \end{aligned}$$

and

$$\widehat{\rho} = \frac{1}{2} \pm \sqrt{\frac{1}{4}+\frac{1}{4+\widehat{\mathbf{V}}^2}} \tag{3}$$

where

$$\begin{aligned} \mathbf{V} =& \frac{p_{123}-p_{12}p_3-p_{13}p_2-p_{23}p_1+2p_1p_2p_3}{\sqrt{(p_{12}-p_1p_2)(p_{13}-p_1p_3)(p_{23}-p_2p_3)}} \\ =& \frac{E\left[\frac{(Y_1-p_1)(Y_2-p_2)(Y_3-p_3)}{p_1\quad p_2\quad p_3}\right]}{\sqrt{E\left[\frac{(Y_1-p_1)(Y_2-p_2)}{p_1\quad p_2}\right]E\left[\frac{(Y_1-p_1)(Y_3-p_3)}{p_1\quad p_3}\right]}E\left[\frac{(Y_2-p_2)(Y_3-p_3)}{p_2\quad p_3}\right]} \end{aligned}$$

For the audiology data, the frequencies in Table 1 yield: $\widehat{p}_1 = 0.000, \widehat{p}_2 = 0.470, \widehat{p}_3 = 0.626, \widehat{p}_{12} = 0.351, \widehat{p}_{13} = 0.423, \widehat{p}_{23} = 0.386, \widehat{p}_{123} = 0.311$. Using these estimates, we arrive at exactly the same values of $(\widehat{\phi}_k, \widehat{\psi}_k),\ k = 1, 2, 3$ and $\widehat{\rho}$ as those calculated earlier by maximizing the likelihood. (Note here that there are two solutions for $\widehat{\rho}$, one larger than $\frac{1}{2}$ and the other smaller. We choose the one that maximizes the likelihood $L = \prod_{i=1}^{n} \rho\phi_1^{Y_1}\phi_2^{Y_2}\phi_3^{Y_3}(1-$

$\phi_1)^{1-Y_1}(1-\phi_2)^{1-Y_2}(1-\phi_3)^{1-Y_3}+(1-\rho)\psi_1^{Y_1}\psi_2^{Y_2}\psi_3^{Y_3}(1-\psi_1)^{1-Y_1}(1-\psi_2)^{1-Y_2}(1-\psi_3)^{1-Y_3}.)$

The likelihood was maximized with a Newton-Raphson scheme using an available Fortran program. One advantage of having the analytic expressions is that estimates can now be calculated directly (even with a hand calculator!) without requiring a numerical optimization routine. More importantly, they describe how relationships observed in the raw data are used to infer properties of the three tests and the prevalence of the latent condition.

In particular, the analytic expression (3) for the estimated prevalence is interesting and novel. It reveals that the starting point for estimation is $\widehat{\rho} = \frac{1}{2}$, with $\mathbf{V}$ determining deviations of $\widehat{\rho}$ from .5; larger values of $\mathbf{V}$ result in lower estimates of ρ. The factor $\mathbf{V}$ compares the three-way association amongst tests in its numerator with the pairwise associations in its denominator. These authors do not yet have an intuitive explanation as to why prevalence is simply a function of the three- versus two-way association parameter under the CI LCA model. It is particularly intriguing that the marginal frequencies of positive tests, p_k, do not directly affect the prevalence estimate. These affect only the true- and false-positive rate estimates (see below). The prevalence estimate is invariant to changes in values of (p_1, p_2, p_3) as long as the three- versus two-way association parameter, $\mathbf{V}$, remains the same.

Somewhat more intuition can be provided for the test accuracy estimates, $\widehat{\phi}_k$ and $\widehat{\psi}_k$, given $\widehat{\rho}$. Consider (1), the estimated sensitivity of the k^{th} test. Note that for a completely uninformative test that has no association with disease status, $P[Y_k = 1|D = 1] = P[Y_k = 1] = p_k$. Thus the starting point for $\widehat{\phi}_k$ is $\widehat{p}_k$, the true-positive rate estimate for an uninformative test. The factor $\widehat{C}_k$, determined by the marginal positive associations between pairs of tests, increases $\widehat{\phi}_k$ above $\widehat{p}_k$. This is logical, since the CI model asserts that any correlation between test results is due to their common association with the latent variable D. If two tests are strongly associated, it must be because they are both accurately reflecting D. The factor C_k is curious in that its numerator reflects associations between Y_k and the other two tests, and its denominator reflects the association between the other two tests. This implies that associations between the k^{th} test and other tests are calibrated by the observed association between those other tests.

The estimates of ϕ_k and ψ_k are very closely linked, since they are determined by exactly the same entities, $\widehat{p}_k, \widehat{C}_k$ and $\widehat{\rho}$. Observe in equations (1) and (2) that if $\widehat{C}_k$ is large, the k^{th} test will be estimated to have a high true-positive *and* a low false-positive rate relative to the uninformative test.

In fact, there is a direct linear relationship between $\widehat{\phi}_k$ and $\widehat{\psi}_k$:

$$\widehat{p}_k = \widehat{\rho}\widehat{\phi}_k + (1 - \widehat{\rho})\widehat{\psi}_k \ .$$

Therefore, given values for $\widehat{\rho}$ and the observed frequency of positive tests, $\widehat{p}_k$, higher estimates of sensitivity also give rise to higher estimates of specificity.

Under the CI LCA model, $(\widehat{\phi}_k, \widehat{\psi}_k, \widehat{\rho})$ are maximum likelihood estimators, and hence are consistent and efficient. Moreover, they seem to represent meaningful quantities. Consider the estimate of ϕ_k. Suppose that, in truth, two of the tests, Y_1 and Y_2, have high true-positive rates, and Y_3 does not. In the observed data, we would expect only weak associations between Y_1 and Y_3 and between Y_2 and Y_3, but a strong association between Y_1 and Y_2. Correspondingly, the C_k factor will be low for $k = 3$, because the numerator is small and the denominator is large. On the other hand, for $k = 1$ (or 2), the denominator and one component of the numerator will be small, canceling each other out to some extent, and C_k will be large due to the strong association between Y_1 and Y_2 in the numerator. Thus, $\widehat{\phi}_1$ (and $\widehat{\phi}_2$) will be large, and $\widehat{\phi}_3$ will be small, as they should be. A similar exercise can be undertaken for the case where two tests, Y_1 and Y_2, have low true-positive rates but Y_3 has a high true-positive rate. Compared to associations between Y_1 and Y_3 and between Y_2 and Y_3, the association between Y_1 and Y_2 will be very weak. This yields a high value of C_3, and hence increases $\widehat{\phi}_3$ well above the starting point $\widehat{p}_3$. On the other hand, C_1 and C_2 will be dominated by the weak association between Y_1 and Y_2, assuming that associations between Y_1 and Y_3 and between Y_2 and Y_3 are of comparable size. Hence, $\widehat{\phi}_1$ and $\widehat{\phi}_2$ will be low. We see once again that the LCA estimates make intuitive sense.

The above discussion focused on $\widehat{\phi}_k$, but analagous considerations hold for $\widehat{\psi}_k$. The starting point for estimating ψ_k is $\widehat{p}_k$, the false-positive rate of the uninformative test. Positive associations between tests in the observed data reduce estimates of ψ_k from this starting point.

In contrast, the value of the estimators, $(\widehat{\phi}_k, \widehat{\psi}_k, \widehat{p})$, when the CI LCA model does not hold is questionable. Although the analytic expressions above now afford them interpretations in terms of the observed data, these do not seem to be generally clinically meaningful entities. Suppose, for example, that there is a latent class, D, but that two tests, say Y_1 and Y_2, are conditionally positively *dependent*. The expressions for $\widehat{\phi}_k$ and $\widehat{\psi}_k$ suggest that the estimates will be biased towards optimistic values. Observed correlation between Y_1 and Y_2 will be stronger than is due simply to D, suggesting that $\widehat{\phi}_k$ will be biased large and $\widehat{\psi}_k$ will be biased small. Indeed, this corroborates the simulation results of Torrance-Rynard and Walter (1997).

4. PARAMETER INTERPRETATIONS VIA THE EM-ALGORITHM

The EM-algorithm is a numerical procedure that allows one to calculate maximum likelihood estimates. In this section, we use the EM-algorithm to derive some interesting alternative expressions for the parameter estimates $(\widehat{\rho}, \widehat{\phi}_K, \widehat{\psi}_K)$.

If the latent variable D were observed, the log-likelihood for the data from the i^{th} subject, $Y_i = \{Y_{i1}, \ldots, Y_{iK}\}$, could be written as

$$\log \mathcal{L}_i^C(\rho, \theta) = D_i \log \rho P_\theta(Y_i|D_i = 1) + (1 - D_i) \log(1 - \rho) P_\theta(Y_i|D_i = 0) \ .$$

Given values for $\rho = \rho^*$ and $\theta = \theta^*$, the expected log-likelihood is

$$\begin{aligned} E_{\rho^*,\theta^*}(\rho, \theta) &= \sum_{i=1}^n E\left\{\log \mathcal{L}_i^C(\rho, \theta)|Y_i\right\} \\ &= \sum_{i=1}^n [P(D_i = 1|Y_i, \rho^*, \theta^*)\left\{\log \rho + \log P_\theta(Y_i|D_i = 1)\right\} \\ &\quad + P(D_i = 0|Y_i, \rho^*, \theta^*)\left\{\log(1 - \rho) + \log P_\theta(Y_i|D_i = 0)\right\}] \qquad (4) \end{aligned}$$

where

$$P(D_i = 1|Y_i, \rho^*, \theta^*) = \frac{P_{\theta^*}(Y_i|D_i = 1)\rho^*}{P_{\theta^*}(Y_i|D_i = 1)\rho^* + P_{\theta^*}(Y_i|D_i = 0)(1 - \rho^*)} \ . \qquad (5)$$

The EM-algorithm proceeds by iteratively maximizing $E_{\rho^*,\theta^*}(\rho, \theta)$ with respect to ρ and θ, and substituting these values for ρ^* and θ^* in the next iteration. The algorithm is completed when (ρ^*, θ^*) have converged. The value of ρ that maximizes (4) is

$$\rho = \sum_{i=1}^{n} P(D_i = 1|Y_i, \rho^*, \theta^*)/n.$$

Therefore, at convergence of the algorithm,

$$\widehat{\rho} = \sum_{i=1}^{n} \widehat{P}(D_i = 1|Y_i)/n, \tag{6}$$

where $\widehat{P}(D_i = 1|Y_i) = P(D_i = 1|Y_i, \widehat{\rho}, \widehat{\theta})$ is given by (5).

The discussion thus far in this section is general in regards to the LCA model, $P_\theta(Y_i|D_i)$. Adding the CI assumption and the notation $\phi = \{\phi_1, \ldots \phi_K\}$ and $\psi = \{\psi_1 \ldots \psi_K\}$ yields the following expression for the expected log-likelihood:

$$\begin{aligned} E_{\rho^*,\theta^*}(\rho, \theta) = \sum_{k=1}^{K}[\sum_{i=1}^{n} & P(D_i = 1|Y_i, \rho^*, \phi^*, \psi^*)\{Y_{ik} \log \phi_k + (1 - Y_{ik}) \log(1 - \phi_k)\} \\ & + P(D_i = 0|Y_i, \rho^*, \phi^*, \psi^*)\{Y_{ik} \log \psi_k + (1 - Y_{ik}) \log(1 - \psi_k)\}] \\ & + \log \rho \sum_{i=1}^{n} P(D_i = 1|Y_i, \rho^*, \phi^*, \psi^*) + \log(1 - \rho) \sum_{i=1}^{n} P(D_i = 0|Y_i, \rho^*, \phi^*, \psi^*). \end{aligned}$$

This expression is maximized at

$$\widehat{\phi}_k = \sum_{i=1}^{n} Y_{ik} P(D_i = 1|Y_i, \rho^*, \phi^*, \psi^*)/\sum_{i=1}^{n} P(D_i = 1|Y_i, \rho^*, \psi^*)$$

and

$$\widehat{\psi}_k = \sum_{i=1}^{n} Y_{ik}\widehat{P}(D_i = 0|Y_i, \rho^*, \phi^*, \psi^*) / \sum_{i=1}^{n} P(D_i = 0|y_i, \rho^*, \phi^*, \psi^*).$$

Therefore, at convergence, the maximum likelihood estimates can be written as

$$\widehat{\phi}_k = \sum_{i=1}^{n} Y_{ik}\widehat{P}(D_i = 1|Y_i)/n\widehat{\rho} \tag{7}$$

$$\widehat{\psi}_k = \sum_{i=1}^{n} Y_{ik}\widehat{P}(D_i = 0|Y_i)/n(1 - \widehat{\rho}). \tag{8}$$

A few observations are warranted at this point. First, expressions (6), (7), (8) do not provide explicit formulas for calculating $\widehat{\rho}$, $\widehat{\phi}$ and $\widehat{\psi}$. Rather they describe some *relationships* among the estimators. Each expression on the right hand side is a function of all three parameters through the terms $\widehat{P}(D_i = 1|Y_i)$. Second, the expressions are intuitive, in the sense that, if $\widehat{P}(D_i = 1|Y_i)$ is an unbiased estimate of $P(D_i = 1|Y_i)$, then $E(\widehat{\rho}) = \rho, E(\widehat{\phi}_k) = \phi_k$ and $E(\widehat{\psi}_k) = \psi_k$. Even if the CI assumption does not hold, the estimators of ρ, ϕ_k and ψ_k are valid as long as $\widehat{P}(D_i = 1|Y_i)$ is valid. We can think of these as the naive estimators when D_i is observed, and, when D_i is not observed, D_i is replaced with $\widehat{P}(D_i = 1|Y_i)$.

One avenue, therefore, for exploring bias in the estimators $\widehat{\rho}, \widehat{\phi}_k$ and $\widehat{\psi}_k$ when CI fails is to consider how violations of the CI assumption affect $\widehat{P}(D_i =$

$1|Y_i)$. For example, in the case of extreme positive dependence between the three tests, i.e., $Y_{i1} = Y_{i2} = Y_{i3}$ almost surely, we would anticipate that $\widehat{P}(D_i = 1|Y_i)$ will be biased large if $(Y_{i1}, Y_{i2}, Y_{i3}) = (1, 1, 1)$ and biased small if $(Y_{i1}, Y_{i2}, Y_{i3}) = (0, 0, 0)$. Expressions (7) and (8) then imply over-optimistic values for $(\widehat{\phi}_1, \widehat{\phi}_2, \widehat{\phi}_3, \widehat{\psi}_1, \widehat{\psi}_2, \widehat{\psi}_3)$ even if $\widehat{\rho}$, the average probability $\sum \widehat{P}(D_i = 1|Y_i)/n$, is unbiased.

In the audiology data, we do in fact have a gold standard measure of disease status. Hence, we can actually compare the observed (true) and latent class estimates of ρ, ϕ_k, and ψ_k. In Table 2 we show the subject-specific estimates of $P(D_i = 1|Y_{i1}, Y_{i2}, Y_{i3})$ for these data. Observe that expressions (6), (7) and (8) do indeed yield the LCA maximum likelihood estimates of ρ, ϕ_k and ψ_k given in Table 1. With D observed, prevalence is calculated as 42%, whereas the LCA estimate that ignores D is 54%.

With data on D available, we can test if the CI assumption holds. A log-linear model yields a likelihood ratio test statistic with three degrees of freedom in both cases and controls. The sum of these two statistics is 169.4 with six degrees of freedom ($p < .001$). Thus, CI does not hold. There is in fact positive dependence amongst tests. As mentioned above, this inflates LCA estimates of $P(D_i = 1|Y_i)$ for subjects with positive tests and deflates LCA estimates of $P(D_i = 1|Y_i)$ for subjects with negative tests. The last two columns of Table 2 bear this out. Correspondingly, the LCA estimates of (ϕ_k, ψ_k) are seen to be over-optimistic relative to their true values calculated using D. We write the true values as

$$\widehat{\widehat{\phi}}_k = \sum Y_{ik} D_i / \sum D_i$$
$$\widehat{\widehat{\psi}}_k = \sum Y_{ik}(1 - D_i) / \sum (1 - D_i)$$

and obtain

$$\widehat{\widehat{\phi}}_1 = 0.664, \widehat{\widehat{\phi}}_2 = 0.625, \widehat{\widehat{\phi}}_3 = 0.751, \widehat{\widehat{\psi}}_1 = 0.401, \widehat{\widehat{\psi}}_2 = 0.360, \widehat{\widehat{\psi}}_3 = 0.537.$$

Table 2 here

Another use of latent class analysis is to derive an operational definition of disease based on observable test results. In this data, we note that the estimates of $P(D_i = 1|Y_{i1}, Y_{i2}, Y_{i3})$ are high for certain combinations of test results, and low for others. In particular, if two or more test results are positive, $\widehat{P}(D_i = 1|Y_i) \geq .78$. On the other hand if two or more are negative, $\widehat{P}(D_i = 1|Y_i) \leq .24$. This result suggests the classification rule that the condition is considered present (absent) if two or more of the tests are positive (negative). However, comparison with the observed D indicates that this LCA based classifier is very poor, with a false-positive rate of 42% and a false-negative rate of 30%. Again, violation of the CI assumption leads to misleading inference.

5. DISCUSSION

Imperfect reference tests are a common problem in the evaluation of diagnostic and prognostic classifiers. Latent class analysis has been promoted heavily in the statistical literature as providing a solution. However, the methodology is not transparent even to those of us who have the highest levels of training in biostatistics. The contribution of this paper is to further our understanding of this popular but technical methodology.

In the special case of three tests where conditional independence is assumed to hold, we derived closed form analytic expressions for maximum likelihood estimates of prevalence and of the associations between observed and latent variables. We found that, given an estimate of prevalence, estimation of the true- and false-positive rates depend on the observed pairwise associations between tests, and on the marginal frequencies of positive tests. This seems intuitive. Less intuitive is the result that the prevalence estimate is a function of the three- versus two-way associations between test results. The unintuitive nature of this estimator reinforces the fact that the estimates have no clinically relevant interpretation and are valid only when the conditional independence latent class model holds.

We have also derived a second set of expressions which describe relationships amongst parameter estimates, and show how they rely on the estimated probabilities that individuals have the condition given observed data. These expressions also make apparent the bias in the paramter estimates when the conditional independence assumption fails.

The analytic results in this paper pertain to the case where three tests are available. With only three tests, a latent variable structure based on conditional independence must be assumed in order to ensure parameter identifiability. The expressions we derived for parameter estimates, however, indicate that they have no merit more generally, i.e., outside of the conditional independence model. This, along with the fact that CI cannot be tested leads us to caution strongly against the use of latent class analysis in practice when only three tests are available.

This argument can be extrapolated to settings with more than three tests. Regardless of how many tests are available, some untestable assumptions must be made for identifiability of test performance parameters. Given our observations, we expect that the estimates obtained will be reasonable only within the context of the assumed (untestable) model, and will have no basis more broadly. This is corroborated by work by Albert and Dodd (2004), who found a very varied set of test performance estimates when they fit different latent class conditional dependence models to the same data. Assumptions about the latent structure impact heavily on inference about test performance, and since the latent structure is unknowable, one cannot endorse the results of latent class analysis as being scientific.

REFERENCES

Albert PS, Dodd LE. (2004) A cautionary note on robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* **60**:427–435.

Albert PA, McShane LM, Shih JH, The US NCI 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.

Boelaert M, Rijal S, Regmi S, Singh R, Karki B, Jacquet D, Chappuis F, Campino L, Desjeux P, Le Ray D, Koirala S, Van der Stuyft P. (2004) A comparative study of the effectiveness of diagnostic tests for visceral leishmaniasis. *American Journal Tropical Medicine Hygeine* **70**(1):72–77.

Dawid AP, Skene AM. (1979) Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics* **28**:20–28.

Dendukuri N, Joseph L. (2001) Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* **57**:158–167.

Espeland MA, Handelman SL. (1989) Using latent class models to characterize and assess relative-error in discrete measurements. *Biometrics* **45**:587–599.

Ferraz MB, Walter SD, Heymann R, Atra E. (1995) Sensitivity and specificity of different diagnostic criteria for Behcet's disease according to the latent class approach. *British Journal of Rheumatology* **34**(10):932-5.

Hui SL, Zhou XH. (1998) Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research* **7**:354–70.

Moayyedi P, Duffy J, Delaney B. (2004) New approaches to enhance the accuracy of the diagnosis of reflux disease. *Gut* **53**(4):iv55–57.

Pepe MS. (2003) The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press.

Pepe MS, Alonzo TA. (2001) Comparing disease screening tests when true disease status is ascertained only for screen positives. *Biostatistics* **2**:249–60.

Qu Y, Tan M, Kutner MH. (1996) Random effects models in latent class analysis for evaluating accuracy of diagnostic test. *Biometrics* **52**: 797–810.

Torrance-Rynard VI, Walter SD. (1997) Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine* **16**:2157–2175.

Walter SD, Irwig LM. (1988) Estimation of test error rates, disease prevalences, and relative risk from misclassified data: A review. *Journal of Clinical Epidemiology* **41**:923-37.

Yang I, Becker MP. (1997) Latent variable modeling of diagnostic accuracy. *Biometrics* **53**: 948–58.

Young J, Bucher H, Tschudi P, Periat P, Hugenschmidt C, Welge-Lussen A. (2003) The clinical diagnosis of acute bacterial rhinosinusitis in general practice and its therapeutic consequences. *Journal Clinical Epidemiology* **56**(4):377–384.

APPENDIX

Derivation of Maximum Likelihood Estimators

We show that under the conditional independence model there is a one-to-one mapping of the 7 parameters $(\rho, \theta) = (\rho, \{(\phi_k, \psi_k), k = 1, 2, 3\})$ to the 7 probabilities $P = (p_1, p_2, p_3, p_{12}, p_{13}, p_{23}, p_{123})$ that characterize the probability distribution of the observable data, i.e., the $2 \times 2 \times 2$ frequency table (e.g., Table 1). Writing this mapping as $g : g(P) = (\rho, \theta)$ and noting that the maximum likelihood estimates of the observable probabilities are the corresponding data frequencies $\widehat{P} = (\widehat{p}_1, \widehat{p}_2, \widehat{p}_3, \widehat{p}_{12}, \widehat{p}_{13}, \widehat{p}_{23}, \widehat{p}_{123})$, it follows that the maximum likelihood estimates of (ρ, θ) are $g(\widehat{P})$.

The following 7 equations follow from elementary probability theory and the conditional independence assumption:

$$p_k = \rho\phi_k + (1-\rho)\psi_k, \quad k = 1, 2, 3 \tag{1a}$$

$$p_{kj} = \rho\phi_k\phi_j + (1-\rho)\psi_k\psi_j, \quad k < j,\ (k, j) \in (1, 2, 3) \tag{2a}$$

$$p_{123} = \rho\phi_1\phi_2\phi_3 + (1-\rho)\psi_1\psi_2\psi_3 \tag{3a}$$

These define g^{-1}. Algebraic manipulations yield the expressions for $(\rho, \{(\phi_k, \psi_k),\ k = 1, 2, 3\})$ in terms of P, i.e., the function g. First we write ψ_k in terms of

(P, ρ, ϕ_k) using (1a)

$$\psi_k = (p_k - \rho\phi_k)/1 - \rho, \quad k \in (1, 2, 3) \tag{4a}$$

and substitute into (2a) to yield

$$(p_k - \phi_k)(p_j - \phi_j) = \frac{1 - \rho}{\rho}(p_{kj} - p_k p_j), \quad k < j, \ (k, j) \in (1, 2, 3).$$

Thus we can write ϕ_2 and ϕ_3 in terms of (P, ϕ_1, ρ):

$$\phi_k = p_k - \frac{(1 - \rho)}{\rho} \frac{p_{1k} - p_1 p_k}{p_1 - \phi_1}, \quad k = 2, 3$$

and substituting into the above expression for $(p_2 - \phi_2)(p_3 - \phi_3)$ we have

$$\begin{aligned}(p_1 - \phi_1)^2 &= \frac{(p_{12} - p_1 p_2)(p_{13} - p_1 p_3)}{p_{23} - p_2 p_3} \frac{(1 - \rho)}{\rho} \\ &= C_1(1 - \rho)/\rho,\end{aligned}$$

where C_1 was defined in Section 3. There are two solutions then for ϕ_1 : $p_1 \pm \sqrt{C_1}\sqrt{(1 - \rho)/\rho}$. We choose $p_1 + \sqrt{C_1}\sqrt{(1 - \rho)/\rho)}$ which follows from the reasonable assumption that the true-positive rate is at least as large as the false-positive rate, $\phi_1 \geq \psi_1$. Similar steps yield $\phi_2 = p_2 + \sqrt{C_2}\sqrt{(1 - \rho)/\rho}$ and $\phi_3 = p_3 + \sqrt{C_3}\sqrt{(1 - \rho)/\rho}$. Substituting ϕ_k into equation (4a) above yields

$$\psi_k = (p_k - \rho p_k - \sqrt{C_k}\sqrt{(1-\rho)\rho})/(1-\rho)$$

$$= p_k - \sqrt{C_k}\sqrt{\rho/(1-\rho)}.$$

Substituting expressions for ϕ_k and ψ_k into equation (3a) and gathering terms yields:

$$p_{123} = p_1p_2p_3\left[1 + \sqrt{C_1C_2} + \sqrt{C_2C_3} + \sqrt{C_1C_3} + \sqrt{C_1C_2C_3}\left\{\sqrt{\frac{1-\rho}{\rho}} - \sqrt{\frac{\rho}{1-\rho}}\right\}\right].$$

Equivalently,

$$\left\{\sqrt{\frac{1-\rho}{\rho}} - \sqrt{\frac{\rho}{1-\rho}}\right\} = \frac{p_{123} - p_1p_2p_3(1+\sqrt{c_1c_2}+\sqrt{c_1c_3}+\sqrt{c_2c_3})}{p_1p_2p_3\sqrt{c_1c_2c_3}},$$

which is easily shown to equal $\mathbf{V}$ as defined in Section 3. Hence,

$$\mathbf{V} = \left\{\sqrt{\frac{1-\rho}{\rho}} - \sqrt{\frac{\rho}{1-\rho}}\right\},$$

and thus

$$\rho = \left\{1 \pm \sqrt{1 + 4/(4+\mathbf{V}^2)}\right\}/2.$$

Table 1:

A. Results of three tests for hearing impairment performed on $n = 666$ subjects.

	$Y_2 = 0$		$Y_2 = 1$	
	$Y_3 = 0$	$Y_3 = 1$	$Y_3 = 0$	$Y_3 = 1$
$Y_1 = 0$	162	85	29	50
$Y_1 = 1$	31	75	27	207

B. Parameter estimates from LCA

$$\begin{aligned}
\widehat{\rho} &= 0.536 \\
(\widehat{\phi}_1, \widehat{\psi}_1) &= (0.841,\ 0.129) \\
(\widehat{\phi}_2, \widehat{\psi}_2) &= (0.762,\ 0.133) \\
(\widehat{\phi}_3, \widehat{\psi}_3) &= (0.898,\ 0.312)
\end{aligned}$$

Table 2: Estimated probabilities of disease $\widehat{P}(D = 1|Y)$ by categories of test results, using LCA (that ignores D) and using the empirical proportions.

$(Y_1,$	$Y_2,$	$Y_3)$	#observations	LCA estimate	Proportion $(D = 1)$
(0,	0,	0)	162	0.0085	0.2346
0	0	1	85	0.1430	0.3176
0	1	0	29	0.1525	0.2069
1	0	0	31	0.2360	0.3548
0	1	1	50	0.7771	0.4400
1	0	1	75	0.8568	0.3733
1	1	0	27	0.8658	0.5185
1	1	1	207	0.9921	0.6329