Weak Identifiability in Latent Class Analysis

Marcus Berzofsky[1], Paul P. Biemer[2],
[1]RTI International, 3040 Cornwallis Rd. RTP, NC 27709
[2] RTI International, 3040 Cornwallis Rd. RTP, NC 27709

Abstract

Model identifiability requires a model likelihood with a single global maximum. The ability to find unique parameter estimates for latent class models depends on such identifiability. Weak identifiability occurs when there are regions of the likelihood that are "flat." In that case, a unique global maximum may exist, but so do many local maxima that provide nearly the same value of the likelihood. In this case, it can be very difficult to find the MLEs, which can lead to erroneous model estimates and conclusions. An important cause of weak identifiability is a violation in one of the model assumptions (e.g., local dependence). This research assesses the likelihood of having a weakly identifiable model based on the type and severity of assumption violation using a simulation approach. It also provides suggestions on how to detect weak identifiability and offers some approaches for avoiding local optima in these situations.

Key Words: Latent class analysis (LCA), weak identifiability, Maximum likelihood estimation

1. Introduction

Latent class analysis (LCA) is a modeling technique used by survey methodologists that utilizes a maximum likelihood estimation process to estimate the level of measurement error in a survey estimate. LCA is a powerful tool in assessing measurement error because it does not require a gold standard (i.e., error free) measurement in order to estimate the measurement error. For dichotomous outcomes, LCA uses a set of indicators to estimate a pre-defined latent variable and estimate the amount of measurement error in those indicators. For example, questions about past year marijuana use are used to estimate the true marijuana use status of a person in the past year. An indicator is a survey item that is highly correlated to the latent construct of interest. A latent variable is an unobserved outcome whose status can only be measured with some level of error.

For dichotomous outcomes, which will be the focus of this paper, LCA uses the indicators and the latent variable to estimate the false negative rate and the false positive rate. The false negative rate is the probability that a respondent provides a negative response when his true status is positive. The false positive rate is the probability that a respondent provides a positive response when his true status is negative. This paper will focus on sensitive outcomes which are outcomes that have a high probability of a respondent providing a false negative answer, but a small or negligible probability of a respondent providing a false positive answer. Examples of sensitive outcomes include drug use and rape or sexual assault.

LCA requires at least three indicators in order to fit. These indicators can come from several different sources. For example, embedded replication can be used which incorporates multiple items within the survey instrument that measure the same latent construct. Other types of indicators can be administrative records or a reinterview measurement. Combinations of these sources can be used in LCA. For instance, a survey with two embedded replications along with an administrative record or reinterview response can be used to obtain the necessary three indicators.

1.1 Identifiability Spectrum

As a model, LCA has certain requirements in order for its estimates to be valid. One key requirement is that the model be identifiable. Our research has found that there is actually a spectrum of four levels of identifiability rather than a simple dichotomy of an identifiable or non-identifiable model. These levels are illustrated in Figure 1 and include:

- *Identifiable model* – a model who has one and only one maximum (Biemer, 2011). Alternatively, this is defined as a model who information matrix is positive definite (i.e., all eigenvalues are positive).
- *Local maxima model* – a model where, upon many iterative runs, many different solutions are obtained. This model type has only one global maximum, but it may be difficult to obtain or distinguish between many similar local maxima.
- *Weakly identifiable model* – a model whose likelihood is fairly flat in the neighborhood of an estimate (Bartholomew and Knott, 1999). This indicates that one or more eigenvaules are positive, but near zero. This may yield two solutions that are close, but still different, coming from the same value of the maximum likelihood.
- *Nonidentifiable model* – occurs when two or more solutions give the same value of the global maximum, but the estimates are not similar.

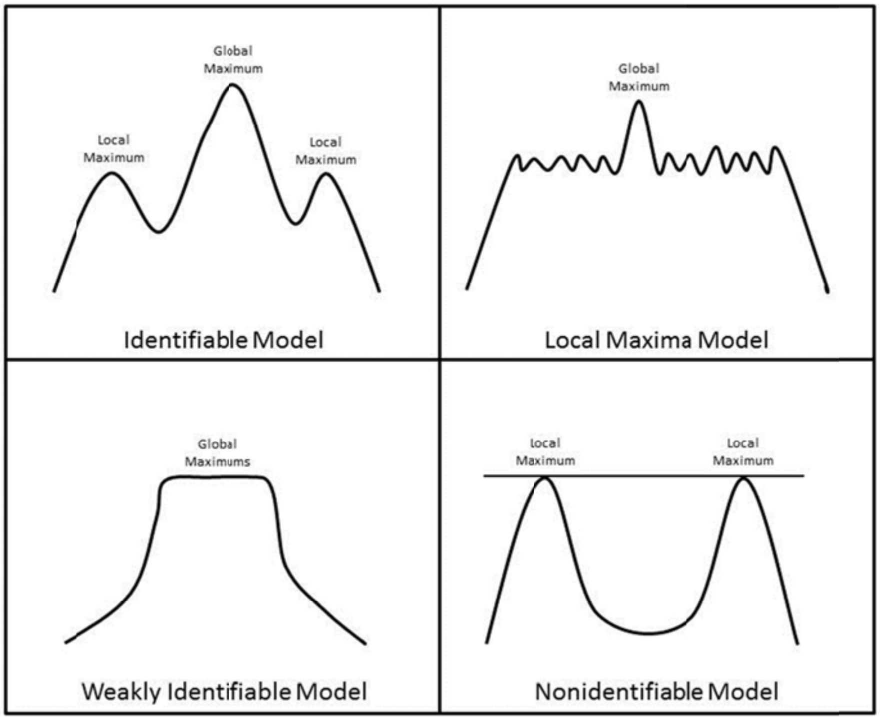

Figure 1. Identifiability spectrum

An understanding of the identifiability spectrum is important for several reasons. For instance, a weakly identifiable model is often an indicator of a large variance. This may lead to unstable estimates (i.e., obtain different estimates through maximum likelihood for the same global maximum). Furthermore, as the likelihood approaches total flatness the model becomes nonidentifiable. Moreover, in the case of a local maxima model, one may not be sure that the global maximum was obtained if the model was only run once. If any of these issues occur they could lead to erroneous or misleading estimates and conclusions.

1.2 Background on the latent class model (LCM)

The LCM was developed by Lazersfeld and Henry (1968). It consists of two components: the structural component and the measurement component. The structural component estimates the latent variable. The measurement component measures the false negative rate and false positive rate of each indicator. The likelihood kernel for the LCM can be written as

$$\pi_{gabc}^{GABC} = \sum_{x} \pi_{g}^{G} \pi_{x|g}^{X|G} \pi_{a|gx}^{A|GX} \pi_{b|gx}^{B|GX} \pi_{c|gx}^{C|GX}$$

Where *X* represents the latent variable with values x=0, 1 for a dichotomous outcome; *A*, *B*, and *C* represent indicators of the latent variable *X* with the same values as *X*; *G* represents a grouping variable with values g=1,2,…,k which is used to create homogeneous groups based on ones probability to provide an erroneous response.

The key assumption in an LCM is local independence (Biemer, 2011). Local independence is defined as

$$\pi_{abc|x}^{ABC|X} = \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{c|x}^{C|X}$$

If local independence is not met the model is locally dependent. Local dependence is caused by one of three sources:

- Bivocality – two or more indicators are measuring different latent variables
- Correlated error – probability of erroneous response dependent on previous response
- Group heterogeneity – error rates differ among respondents in a particular group

As shown in Berzofsky (2011), if one of these sources of error occurs the estimates of measurement error will be biased. The direction and magnitude of the bias depend on the source of error and level to which the error occurs (e.g., if bivocality occurs the less correlated an indicator is to the latent variable the greater the bias).

Another potential issue that one needs to be aware of when conducting a LCM is sparseness of the data. Sparseness occurs when several of the cell sizes in the data frequency table are zero. Depending on the severity, sparseness can sometimes lead to problems with model fit.

2. Study Questions and Methods

2.1 Study Questions

Our study had two questions we wanted to answer.

1. Can one of the sources of local dependence (i.e., bivocality, correlated error, group heterogeneity) cause weak identifiability or local maxima models?
2. Can sparseness of data cause a weak identifiability or a local maxima model?

2.2 Methods

2.2.1 Assessing sources of local dependence

A simulation study was conducted to assess whether one of the sources of local dependence caused weak identifiability or a local maxima model. In this paper we only looked at the sources bivocality and correlated error.

For each source of error, we induced varying levels of failure in the model assumption. As shown in Berzofsky (2011), each source of error can be written in terms of a correlation. For bivocality the correlation between the latent variables being measured indicates the level of bivocality. When there is perfect correlation the two indicators are univocal (i.e., there is no bivocality). However, when the correlation is zero the two indicators are measuring completely different latent variables. For correlated error, the correlation between the two indicators quantifies the dependency of the two indicators. For instance, when the correlation is zero, the two indicators had no dependency which means that the error rates are independent. However, when there is perfect correlation that means that the probability of providing an erroneous response in one indicator is completely dependent on the other indicator.

The simulation was run 1,000 times for each level of correlation. The number of times the global maximum was reached was recorded. When the global maximum was reached it was verified whether the same estimate was produced.

2.2.2 Assessing sparseness in the data

A simulation study was conducted to assess whether sparseness caused weak identifiability or a local maxima model. Using known population parameters – population size, true prevalence rate, false negative rate, and false positive rate – a data frequency table was constructed.

In order to test for sparseness the population size was reduced. For each population size tested the model was run 100 times. For each population size the number of times the global maximum was reached was recorded. When the model converged to the global maximum, it was verified whether the same estimates were produced.

3. Results

3.1 Sources of local dependence

3.1.1 Bivocality

As shown in Figure 2, when bivocality was induced the as the correlation between the two latent variables decreased the number of times the model converged on the global maximum decreased. The rate at which the decreased occurred increased as the prevalence became rarer. As shown in Figure 3, as the false negative rate increased the rate at which the model converged at the global maximum increased. These findings are indicative of a local maxima model. However, in all cases, when the mode converged to the global maximum, the same estimates were produced indicating that weak identifiability did not occur.

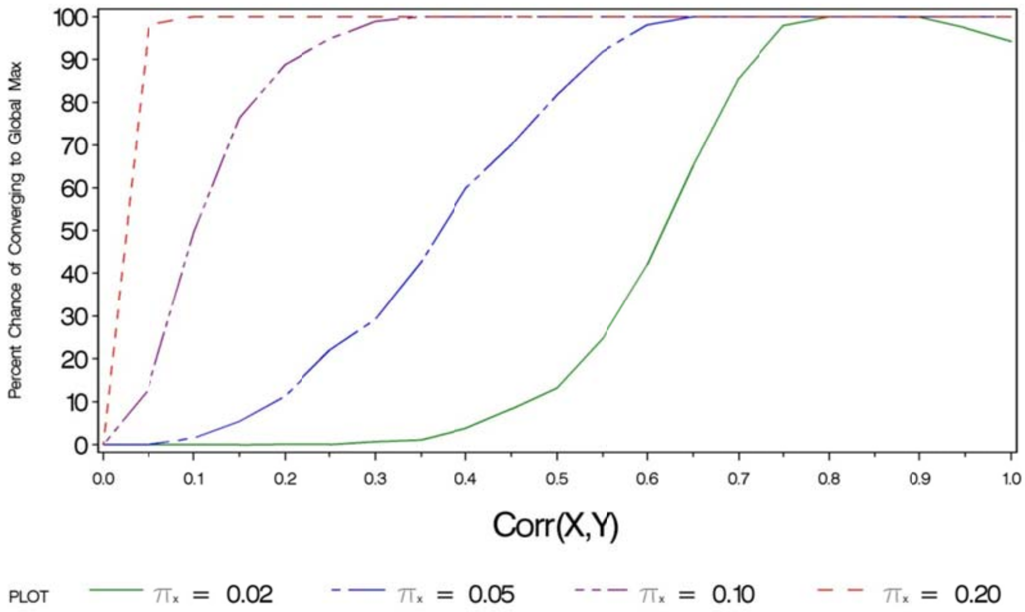

Figure 2. Percent chance of convergence to the global maximum as the correlation between two latent variables decreases (bivocality) by varying levels of true prevalence rate.

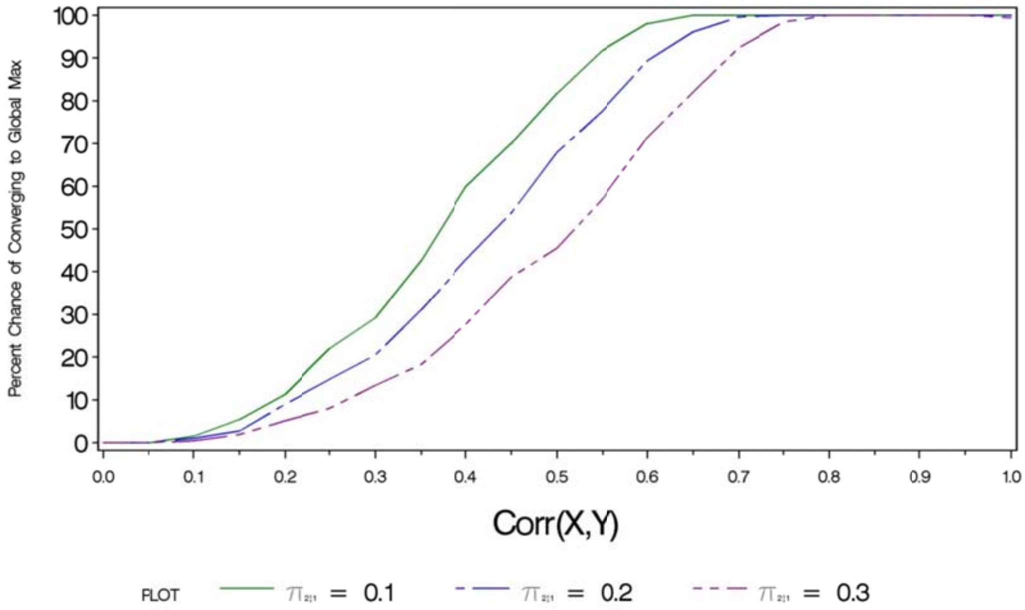

Figure 3. Percent chance of convergence to the global maximum as the correlation between two latent variables decreases (bivocality) by varying levels of the false negative rate.

3.1.2 Correlated error

As shown in Figure 4, as the correlation between the two indicators increases, the rate of convergence to the global maximum accelerates. The acceleration increases as the prevalence rate decreases (i.e., the estimate becomes rarer). Furthermore, as shown in Figure 5, as the prevalence rate decreased the rate by which the model failed to converge to the global maximum increased. These findings indicate that local maxima models occur as the correlation between the two indicators increases. When the model converged to the global maximum the same estimates were reached indicating that weak identifiability did not occur.

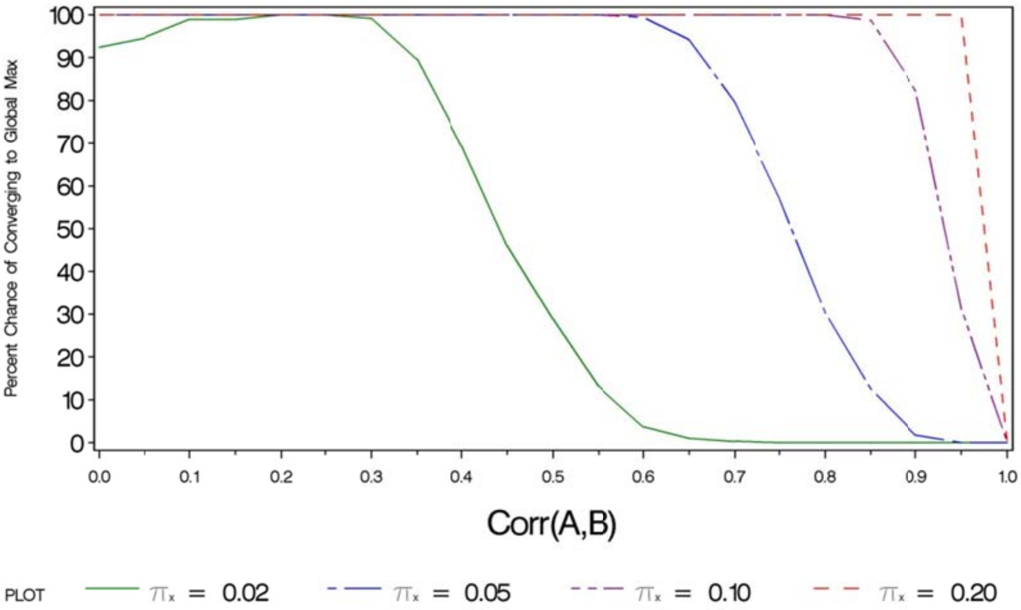

Figure 5. Percent chance of convergence to the global maximum as the correlation between two indicators increases (correlated error) by varying levels of the false negative rate.

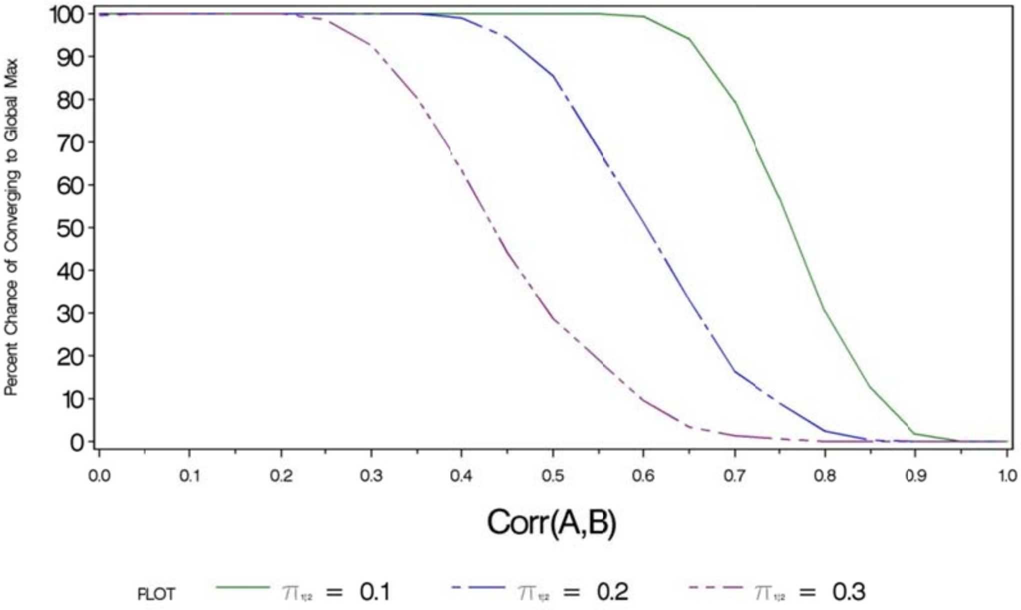

Figure 6. Percent chance of convergence to the global maximum as the correlation between two indicators increases (correlated error) by varying levels the true prevalence.

3.2 Sparseness of data

As shown in Table 1, as the population size decreased the number of empty cells increased. In all cases, convergence to the global maximum occurred indicating that these were not local maxima models. However, When the sample size reached 100 the model became weakly identifiable whereby different estimates were obtained for model iterations with the same maximum likelihood.

Table 1: Percent of empty cells and identifiability of model by population size

Population size	Percent of empty cells	Identifiable model
5,000	0	Yes
4,000	31.25	Yes
3,000	62.5	Yes
2,000	62.5	Yes
1,000	62.5	Yes
500	62.5	Yes
100	78.1	No

4. Discussion

Our study found that failure to meet local dependence leads local maxima models. This makes it more difficult and less likely that the model will converge to the global maxima. Furthermore, our study found that when sparseness of the data increases the model does not have problems converging to the global maximum, but becomes weakly identified.

Given these findings, our research has found two potential options to mitigate these issues. First, in order to increase the success of converging to the global maximum, one can incorporate starting values in the model specifications. Starting values – initial specifications of the parameter values – indicate where the maximum likelihood process should start its search. By indicating values that closely approximate the true classification error probabilities, the chance of convergence to the global maximum is greatly increased. However, it is important to have a good sense of what the parameter values are. Using poor starting values will often not be successful if a model is weakly identified.

Alternatively, if the cause of weak identifiability is local dependence, one can attempt to correct the model assumption failure. For both bivocality and correlated error this can be done by adding direct effects. Direct effects are the interaction between two indicators (Hagenaars, 1988; Berzofsky, 2011). However, when there are only three indicators, additional constraints on the model may be necessary to ensure identifiability. Furthermore, when there are four or more indicators, when there is bivocality, a second latent variable can be added to the model (Berzofsky, 2011).

References

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis*. London: Arnold.

Berzofsky, M.E. (2011). Methods and Approaches for Evaluating the Validity of Latent Class Models with Applications. An unpublished dissertation.

Biemer, Paul P. 2011. *Latent Class Analysis of Survey Error*, John Wiley & Sons, Hoboken, N.J.

Hagenaars, Jacques A. 1988. “Latent Structure Models with Direct Effects between Indicators: Local Dependence Models.” *Sociological Methods and Research* 16:379 – 405.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

.