Power Analysis for the Likelihood-Ratio Test in Latent Markov Models: Short-cutting the bootstrap p-value based method

Dereje W. Gudicha, Verena D. Schmittmann, Fetene B. Tekle, and Jeroen K. Vermunt
Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands

Abstract

The latent Markov (LM) model is a popular method for identifying distinct unobserved states and transitions between these states over time in longitudinally observed responses. The bootstrap likelihood-ratio (BLR) test yields the most rigorous test for determining the number of latent states, yet little is known about power analysis for this test. Power could be computed as the proportion of the bootstrap p-values (PBP) for which the null hypothesis is rejected. This requires performing the full bootstrap procedure for a large number of samples generated from the model under the alternative hypothesis, which is computationally infeasible in most situations. This paper presents a computationally feasible short-cut method for power computation for the BLR test. The short-cut method involves the following simple steps: 1) obtaining the parameters of the model under the null hypothesis, 2) constructing the empirical distributions of the likelihood-ratio under the null and alternative hypotheses via Monte Carlo simulations, and 3) using these empirical distributions to compute the power. We evaluate the performance of the short-cut method by comparing it to the PBP method, and moreover show how the short-cut method can be used for sample size determination.
Keywords: Latent Markov, Number of States, Likelihood-Ratio, Bootstrap, Monte Carlo simulation, Longitudinal Data, Power Analysis, sample size.

1 Introduction

In recent years, the latent Markov (LM) model has proven useful to identify distinct underlying states and the transitions over time between these states in longitudinally observed responses. In LM models, as in latent class models, or more generally in finite mixture models, the observed responses are governed by a set of discrete underlying categories, which are named states, classes, or mixture components. Moreover, the LM model allows transitions between these states from one time-point to another, that is, the state membership of respondents can change during the period of observation. The LM model finds its application, for example, in educational sciences to study how the interests of students in certain subjects changes over time (Vermunt et al., 1999), and in medical sciences to study the change in health behavior of patients suffering from certain diseases (Bartolucci et al., 2010). Various examples of applications in social, behavioral, and health sciences are presented in the textbooks by Bartolucci et al. (2013) and Collins and Lanza (2010).

In most research situations, including those just mentioned, the number of states is unknown and must be inferred from the data itself. The bootstrap likelihood-ratio (BLR) test, proposed by McLachlan (1987) and extended by Feng and McCulloch (1996) and Nylund et al. (2007), is often used to test hypotheses about the number of mixture components. These previous studies focused on p-value computation, rather than on power computation for the BLR test, which is the topic of the current study.

The assessment of the power of a test, that is, the probability that the test will correctly reject the null hypothesis when indeed the alternative hypothesis is true, is important at several stages of a research study. At the planning stage, an a priori power analysis is useful for determining the data requirements of the study: e.g., the sample size, and the number of time points at which measurement takes place. In general, the smaller the sample size, the less power we have to reject the null hypothesis when it is false. Therefore, a too small sample size may result in an under-extraction of the number of states (see for example, Nylund et al. (2007) and Yang (2006)). This not only misleads the conclusion about the number of states but also the interpretation of the state specific parameters. Moreover, when the sample size is too small, the parameter estimates are prone to be unstable and inaccurate (Collins and Wugalter, 1992; Marsh et al., 1998). Performing an a priori power analysis helps to determine the smallest sample required to

achieve a certain power level, usually a power level of .8 or larger, thereby allowing the researcher to avoid excessively large, uneconomical sample sizes. Oftentimes, when applying for a research grant, the funding agency asks to justify the number of subjects to be enrolled for the study through a power analysis. At the analysis stage, a post hoc assessment of the power achieved given the specific design scenario and the parameter values obtained should aid the interpretation of the study results. Also, in order to assure confidence in the study results (or conclusions), journal editors often ask to report the power.

Power computation is straightforward if, under certain regularity conditions, the theoretical distributions of the test statistic under the null and the alternative hypothesis are known. This is not the case for the BLR test in LM models. The power of a statistical test can be computed as the proportion of the p-values smaller than the chosen alpha. When using the BLR statistic to test for the number of states in LM models, such a power calculation becomes computationally expensive, because it requires performing the bootstrap p-value computation for multiple sets of data. As explained in detail below, it requires generating M data sets from the model under the alternative hypothesis, and for each data set, estimating the models under the null and alternative hypotheses to obtain the LR value. Whether the null hypothesis will be rejected for a particular generated data set is determined by computing the bootstrap p-value, which in turn requires (a) generating B data sets from the model estimates under the null hypothesis and (b) estimating the models under the null and alternative hypotheses using these B data sets. Hereafter, we refer to this computationally demanding procedure, which involves calculating the power as the proportion of the bootstrap p-value for which the model under the null hypothesis is rejected, as the PBP method.

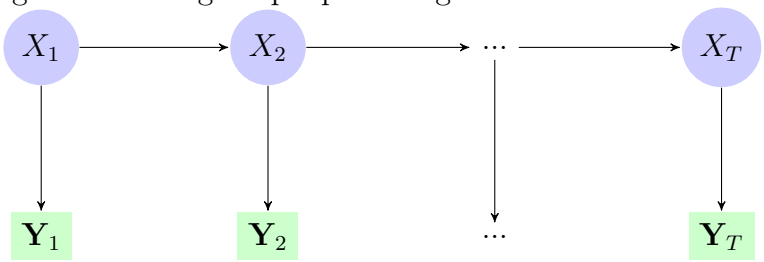
Because using the PBP method is infeasible in most situations, we propose an alternative method which we refer to as the short-cut method. Computing the power using the short-cut method involves constructing the empirical distributions of the LR under both the null and alternative hypotheses. We show how the asymptotic values of the parameters of the model under the null hypothesis can be obtained based on a certain large data set, and these parameters will in turn be used in the process to obtain the distribution of the LR statistic under the null hypothesis. As explained in detail below, the distribution of the LR under the null hypothesis is used to obtain the critical value, given a predetermined level of significance. Given this critical

value, we compute the power by simulating the distribution of the LR under the alternative hypothesis. Using numerical experiments, we examine the data requirements (e.g., the sample size, the number of time points, and the number of response variables) that yield reasonable levels of power for given population characteristics.

The remaining part of the paper is organized as follows. In section 2, we describe the LM model and the BLR test for determining the number of states. In section 3, we provide power computation methods for the BLR test and discuss how these methods can be applied to determine the required sample size. Numerical experiments that illustrate the proposed methods of power and sample size computation are presented in section 4. The paper ends with a discussion and conclusions in section 5.

2 The LM models

Let $\mathbf{Y}_t = (Y_{t1}, Y_{t2}, Y_{t3}, ...Y_{tP})$ for $t = 1, 2, 3, ..., T$ be the P-dimensional response variable of interest at time point t. Denoting the latent variable at time point t by X_t, in a LM model the relationships among the latent and observed response variables at the different time points can be represented using the following simple path diagram.



An LM model is a probabilistic model defining the relationships between the time-specific latent variables X_t (e.g., between X_1, X_2, and X_3) and the relationships between the latent variables X_t and the time-specific vectors of observed responses $\mathbf{Y}_t$ (e.g., X_1 with $\mathbf{Y}_1$). In the basic LM model, the latent variables are assumed to follow a first-order Markov process (i.e., the state membership at t+1 depends only on the state occupied at time point t), and the response variables are assumed to be locally independent given the latent states. Based on these assumptions, we define the S-state LM model as a

mixture density of the form

$$p(\mathbf{y}_i, \Phi) = \sum_{x_1=1}^{S} \sum_{x_2=1}^{S} \sum_{x_3=1}^{S} ... \sum_{x_T=1}^{S} p(x_1) \prod_{t=2}^{T} P(x_t|x_{t-1}) \prod_{j=1}^{P} p(y_{tji}|x_t),$$

where $\mathbf{y}_i$ denotes the vector of responses for subject i over all the time points, y_{tji} the response of subject i to the j-th variable measured at time point t, x_t a particular latent state at time point t, and Φ the vector of model parameters (Vermunt et al., 1999; Bartolucci et al., 2013).

The LM model has three sets of parameters:

1. The initial state probabilities (or proportions) $p(X_1 = s) = \pi_s$ satisfying $\sum_{s=1}^{S} \pi_s = 1$. That is, the probability of being in state s at the first time point;

2. The transition probabilities $p(X_t = s|X_{t-1} = r) = \pi_{s|r}^t$ satisfying $\sum_{s=1}^{S} \pi_{s|r}^t = 1$. These transition probabilities indicate the probabilities of remaining in a state or switching to another state, conditional on the state membership at the previous time point. All transition probabilities are conveniently collected in a transition matrix, in which the entry in row r and column s represents the probability of a transition from state r at time point $(t-1)$ to state s at time point t;

3. The state-specific parameters of the density function $p(y_{tji}|x_t)$, which govern the association between the latent states and the observed response variables. The choice of the specific density form for $p(y_{tji}|x_t)$, which depends on the scale type of the response variable, determines the state-specific parameters for this density function. With continuous responses, one may, for example, define the state-specific density to be a normal distribution, for which the parameters are the mean $\mu_{j|s}^t$ and the variance $\sigma_{j|s}^{2^t}$ (Schmittmann et al., 2005). With dichotomous and nominal responses, the multinomial distribution is assumed, for which the parameters become the conditional response probabilities $p(y_{tji}|x_t = s) = \theta_{j|s}^t$ (Collins and Wugalter, 1992; Vermunt et al., 2008). The state-specific parameters and the transition probabilities may vary across time, hence the subscript t, but are assumed to be time-homogeneous during the remainder of this paper.

Given a sample of size n, the parameters are typically estimated by maximizing the log-likelihood function:

$$l(\Phi) = \sum_{i=1}^{n} \log p(\mathbf{y}_i, \Phi). \tag{1}$$

The search for the values of Φ that maximize the log-likelihood function in equation (1) can be carried out with the Expectation-Maximization (EM) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2007), which alternates between computing the expected complete data log-likelihood function (E step) and updating the unknown parameters of interest by maximizing this function (M step). For LM models, a special version of the EM algorithm with a computationally more efficient implementation of the E step may be used. This algorithm is referred to as the Baum-Welch or forward-backward algorithm (Baum et al., 1970; Bartolucci et al., 2010; Vermunt et al., 2008).

As already discussed in the introduction section, identifying the number of latent states is a common goal in LM modeling, and typically the first step in the analysis. Testing hypotheses about the number of states involves estimating LM models with increasing numbers of states and checking whether the model fit is significantly improved by adding one or more states. More formally, the hypotheses about the number of states may be specified as $H_0 : S = r$ versus $H_1 : S = s$, where $r < s$. Usually, the r-and s-state model differ by one state. For example, the test for H_1 : 3-state LM model against H_0 : 2-state LM model. However, in principle, the comparison can also be between the 3-state and the 1-state LM model. In this paper, we restrict ourselves to the situation in which $r = s - 1$.

The LR statistic for this type of test is defined as

$$LR = 2(l(\hat{\Phi}_s) - l(\hat{\Phi}_r)), \tag{2}$$

where $l(\cdot)$ is the log-likelihood function and $\hat{\Phi}_s$ and $\hat{\Phi}_r$ are the maximum likelihood estimates under the alternative and null hypothesis, respectively. In the standard case, under certain regularity conditions, it is generally assumed that the LR statistic in equation (2) follows a central chi-square under the null hypothesis and a non-central chi-square distribution under the alternative hypothesis (Steiger et al., 1985). In such a case, one may use the (theoretical) chi-square distribution with the appropriate number of degrees of freedom to compute the p-value of the LR test given a predetermined level

of significance α or the power of the LR test given the population characteristics of H_1 model. These asymptotic distributions however do not apply when using the LR statistic for testing the number of latent states (Aitkin et al., 1981). One reason is that the H_0 model with $S-1$ states is obtained from the H_1 model by restricting the initial probability for state S and the transition probabilities towards state S to 0. This violates the regularity condition that restriction should not be on the boundary of the parameter space. In addition, when state S is assumed to have a zero probability of occurrence, the parameters for this state are unidentified, which yields a violation of the regularity condition that all parameters in the H_0 should be identifiable.

One may however apply the method of parametric bootstrapping to construct the empirical distribution of the LR, and subsequently use the contructed empirical distribution for p-value computation. Due to advances in computing facilities, this can be applied readily. Using parametric bootstrapping, the empirical distribution of the LR statistic under the null hypothesis is constructed by generating B independent (bootstrap) samples according to a parametric (probability) model $P(\mathbf{y}, \hat{\Phi}_r)$, where $\hat{\Phi}_r$ itself is an estimate computed based on a sample of size n (McLachlan, 1987; Feng and McCulloch, 1996; Nylund et al., 2007). Denoting the bootstrap samples by $\mathbf{y}^b$ (for $b = 1, 2, 3, ...B$), equation (2) becomes

$$BLR_b = 2(l(\hat{\Phi}_s^b) - l(\hat{\Phi}_r^b)), \tag{3}$$

where BLR_b denotes the BLR, computed for (bootstrap) sample $\mathbf{y}^b$.

So, sampling B data sets from the r-state LM model defined by $P(\mathbf{y}, \hat{\Phi}_r)$ and computing the BLR statistic as shown in equation (3) for each of these data sets, yields the BLR distribution under the null hypothesis. This distribution is then employed in the bootstrap p-value computation. In short, the bootstrap p-value computation proceeds as follows:

Step 1. Treating the ML parameter estimates as if they were the "true" parameter values for the r-state LM model, generate B independent (bootstrap) samples from the r-state LM model.

Step 2. Compute the BLR_b values as shown in equation (3), which requires us to fit the r- and s-state models using the bootstrap samples generated in Step 1.

Step 3. Compute the bootstrap p-value as $p = \frac{1}{B}\sum_{b=1}^{B} I(BLR_b > LR)$, where $I(\cdot)$ is the indicator function which takes on the value 1 if the argument $BLR_b > LR$ holds and 0 otherwise. The decision concerning whether

the r-state LM model should be retained or rejected in favor of the s-state model is then determined by comparing this p-value with the predetermined significance level α.

3 Power analysis for the BLR test

As mentioned, two common goals of power analysis are (a) to determine the post hoc power of a study (i.e., given a certain samples size, number of time points, and number of response variables) and (b) to a priori determine the sample size (or other design factors like the number of time points or the number of response variables) required to achieve a certain power level. In both cases, we assume that the population parameters are known (in a priori analyses a range of expected parameter values may be used) and other factors such as the number of indicator variables and the number of classes are fixed. In what follows, we first show how the bootstrapping procedure discussed above can be used for power computation, and subsequently present the computationally more efficient short-cut method for power and sample size computation in LM models.

3.1 Power computation

In this sub-section, we present two alternative methods for computing the power of the BLR test. The first option, the PBP method, involves computing the power as the proportion of the bootstrap p-values (PBP) for which H_0 is rejected. More specifically, the PBP method for power computation involves the following steps:
Step 1. Generate M independent samples, each of size n, from the parametric model $P(y, \Phi_s)$, where Φ_s is the given parameter values under H_1.
Step 2. For each sample m (m = 1, 2, 3, ..., M) in *Step 1*, compute the likelihood-ratio LR_m as shown in equation (2).
Step 3. Obtain the bootstrap p-value of each sample m as $p_m = \frac{1}{B}\sum_{b=1}^{B} I(BLR_{bm} > LR_m)$, where LR_m is the LR of sample m from the H_1 population, BLR_{bm} is the corresponding BLR for bootstrap sample b, and $I(\cdot)$ is the indicator function as defined above.
Step 4 The actual power associated with a sample of size n is computed as

the proportion of the H_1 data sets in which H_0 is rejected. That is,

$$PBP = \frac{1}{M}\sum_{m=1}^{M} I(p_m < \alpha), \tag{4}$$

where the indicator function $I(\cdot)$ and α are as defined above.

As mentioned above, such a method of power computation is computationally expensive and requires considerable amount of computer memory. For example, setting $M = 500$ and $B = 99$ requires us to generate and analyze $M(B + 1) = 50000$ data sets. Also, in order to achieve a good approximation to the sampling distribution, which, if not well approximated, could affect the p-value (and subsequently the power), both M and B should be large enough.

For LM models, for which model fitting requires iterative procedures, power computation by using the PBP method is computationally too intensive in practice. We propose a computationally more efficient method, which we call the shortcut method. It works very much as the standard power computation (see for example, Brown et al. (1999)), with the difference that we construct the distributions under H_0 and H_1 by Monte Carlo simulation. In Figure 1, these two distributions are indicated with curve H_0 and H_1, respectively. As explained below, the distribution under H_0 is used to obtain the critical value (CV), and the distribution under H_1 is used to compute the power given the CV.

First, the H_0 "population" parameters needed to compute the CV should be obtained. This can be achieved by creating an exemplary data set, which is a data file with all possible response patterns and the relative frequencies of the response patterns under H_1 as weights (O'Brien, 1986; Self et al., 1992). Because in LM models with more than a few indicators and/or time points, the number of possible response pattern is very large, this method cannot always be applied. Therefore, as an alternative, using the parameter values of the H_1 model, we generate a large data set (e.g., 10000 observations), which is assumed to represent the hypothetical H_1 population. Estimating the H_0 model (i.e., the r-state LM model) using this large data set yields the pseudo parameter values for the r-state model. These H_0 parameters are then employed to construct the distribution of the LR under the null hypothesis. That is, given the estimated parameters of the H_0 model, generate K data sets (each of size n) and for each of these data sets, compute the LR as shown in equation (2). Next, order the LR values in such a way that $LR_{[1]} \leq$

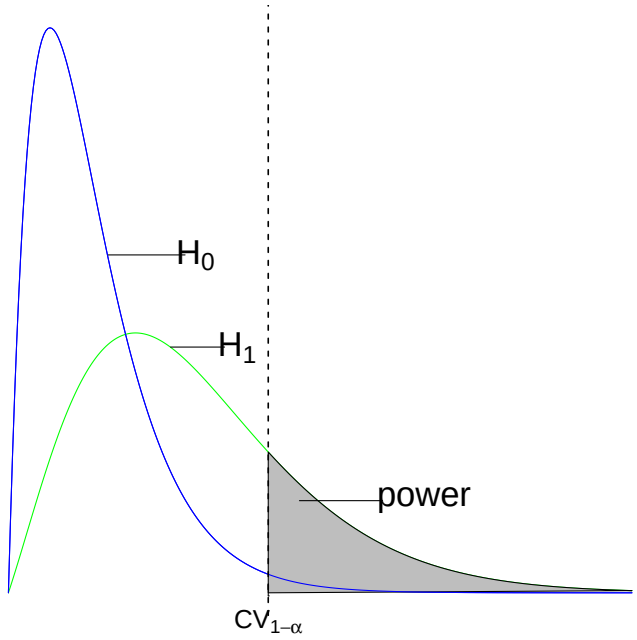


Figure 1: Distributions of LR under the null and alternative hypotheses

$LR_{[2]} \leq LR_{[3]} \leq ... \leq LR_{[K]}$. Given the nominal level α, compute the CV as

$$CV_{(1-\alpha)} = \{LR_k : p(LR > LR_{[k]}|H_0) = \alpha\}. \tag{5}$$

Similarly, the distribution of the LR under the alternative hypothesis is constructed using M samples of the H_1 model. That is, given the parameters of the H_1 model, we generate M independent samples from the s-state LM model and for each of these samples, compute the LR as shown in equation (2). For sufficiently large M, the distribution of the LR under the alternative hypothesis approximates the H_1 curve in Figure 1. The power is then computed as the probability that the LR value belongs to the shaded region of Figure 1. That is,

$$\text{power} = p(LR > CV_{(1-\alpha)}|H_1) = \frac{\sum_{m=1}^{M} I(LR_m > CV_{(1-\alpha)})}{M}, \tag{6}$$

where $I(\cdot)$ is the indicator function, indicating whether the LR value (computed based on the b sample of the H_1 population) exceeds the $CV_{1-\alpha}$ value.

So both, the PBP and the short-cut methods require M samples given H_1 and the calculation of the LR for each of these samples (i.e., steps 1 and 2 of the PBP power calculation). The saving in computation time of the short-cut method lies in the omission of the full bootstrap for each of the M samples from the H_1 model. Rather, the LRs given H_1 are now evaluated against the approximated distribution of LRs given H_0. Therefore, compared to the PBP-based power computation, the number of data sets to be generated and analyzed is much smaller when using the short-cut method. For example, for $M = 500$ and $K = 500$, we analyze $M + K = 1000$ data sets. To further explain the computational time gain, let the time required to calculate the PBP based power by analyzing $M(B+1)$ data sets be ω. The time required to compute the power by the short cut method – which requires analyzing $M + K$ data sets – can be shown to be $\left(\frac{1}{B+1} + \frac{K}{M+\frac{B}{M}}\right)\omega$. For large M, and under the setting with $B = K = M$, this computational time may simplify to $(\frac{2}{M})\omega$. In other words, the shortcut method is $M/2$ times faster than the PBP method.

The short-cut method of power computation presented above can easily be implemented using statistical software for LM analysis as outlined below.

1. Obtain the H_0 population parameters: Given the parameters of the H_1 model, generate a large data (e.g., 10000 observations) from the H_1 population. For this purpose, any software that allows generating a sample from a LM model with fixed parameter values can be used. For the numerical studies shown below, we used the syntax module of the Latent GOLD 5.0 program (Vermunt and Magidson, 2013). Using this large data set, then estimate the parameters of the H_0 model.

2. Compute the CV: Given the estimated parameters of the H_0 model, generate K data sets (each of size n) and for each of these data sets, compute the LR as shown in equation (2). Note that this requires estimating both the r- and the s-state model. For a sufficiently large K, the LR distribution approximates the population distribution of the LR under the null hypothesis (i.e., the H_0 curve in Figure 1). We use this distribution to compute the CV of the LR test as shown in equation (5).

3. Compute the power: Given the parameters of the H_1 model, obtain the empirical distribution of the LR. That is, generate M data sets from

H_1 model, and, using these data sets, compute the LR as shown in (2). Given the CV and the empirical distribution of the LR under H_1, compute the power as shown in equation (6).

3.2 Sample size computation

In this section, we show how the procedure described above for power computation using the short-cut method can be applied for sample size determination. For sample size determination, step 1 of the power computation procedure (discussed under software implementations) remains the same. The last two steps are however repeated for different trial sample sizes. More specifically, suppose the investigator wishes to achieve a certain pre-specified power level (say, power = .8 or larger) while avoiding the sample size to become unnecessarily large. Then, the LR power computation is performed as outlined in step 2 and 3, starting with a certain sample size n_1. Below we provide power curves that can be used as a guidance to locate this starting sample size. If the power obtained based on these n_1 observations is lower than .8, repeat step 2 and 3 by choosing n_2 larger than n_1. If the chosen n_1 result in larger power instead (and we want to optimize the sample size), choose n_2 smaller than n_1 and repeat step 2 and 3. In this way, the power computation procedure is repeated for different trial samples of varying sizes, and from these trial samples, the one that best approximates the desired power level is used as the sample size for the study concerned. In our numerical study, we repeat this power computation procedure for different sample sizes, which resulted in a series of power values. By plotting these power values against the corresponding sample size, we obtain a power curve from which one can easily determine the minimum sample size that satisfies the power requirements, for example that the power should be larger than .8.

When designing a longitudinal study, it is also of interest to determine the number of time points required to achieve a certain power level. For a fixed sample size, a fixed number of response variables, and a priori specified H_1 parameter values, the procedures discussed above for sample size determination can be applied to the number of time points determination as well. More specifically, in step 2 and 3 of the power computation procedures, the number of time points T should be varied instead of the sample size n.

4 Numerical study

A numerical study was conducted to (a) illustrate the proposed power and sample size computation methods, and (b) investigate whether the short-cut method and the PBP method give similar results. This numerical study has an additional benefit for applied researchers using the LM model: given the population characteristics, the resulting BLR power tables and the power curves shown below may help to make an informed decision about the data requirements in testing the number of states for the LM model. More specifically, the results of this numerical study may be used as a guidance by applied researchers to locate the initial trial sample size when computing the required sample size to achieve a desired power level, as discussed in section 3.

4.1 Numerical study set up

The power of the BLR test for the number of states in LM models depends on several design factors and population characteristics. See, for example, Gudicha et al., (2015) who studied factors affecting the power in LM models. The design factors include the sample size, the number of time points, and the number of response variables. The number of latent states, and the various model parameter values (i.e., parameter values for the initial state proportions, for the state transition probabilities, and for the state specific densities) define the population characteristics (Collins and Wugalter, 1992).

In this numerical study, we varied both the design factors and the population characteristics. The design factors varied were the sample size ($n = 300$, 500, or 700), the number of time points ($T = 3$ or 5), and the number of response variables ($P = 6$ or 10). The population characteristics under the alternative hypothesis (i.e, the s-state LM model for $S = 3$, or 4) were specified to meet varying levels of a) initial state proportions (balanced, moderately imbalanced, highly imbalanced), b) stability of state membership (stable, moderately stable, unstable), and c) state-response associations (weak, moderate, strong) as follows.

In line with Dias (2006), the initial state proportions were specified using $\pi_s = \frac{\delta^{s-1}}{\sum_{h=1}^{S} \delta^{h-1}}$. We set the values of δ to $1, 2$, and 3, which correspond to balanced, moderately imbalanced, highly imbalanced initial state proportions, respectively. For the transition matrix, we used the specification suggested by Bacci et al. (2014), which under the assumption of time homogeneity

Table 1: Values of conditional response probabilities

state-responses association levels	S=3			S=4			
	$s=1$	$s=2$	$s=3$	$s=1$	$s=2$	$s=3$	$s=4$
Weak	.75	.58	.25	.75	.58	.75 or.25	.25
Moderate	.80	.65	.20	.80	.65	.80 or .20	.20
Strong	.85	.70	.15	.85	.70	.85 or.15	.15

gives $\pi_{s|r} = \frac{\rho^{|s-r|}}{\sum_{h=1}^{S} \rho^{|h-r|}}$. Setting the values of ρ to $\rho = 0.1, 0.15$, and 0.3 yields what we referred to above as stable, moderately stable, and unstable state membership. In this numerical study, we restricted ourselves to the situation that the response variables of interest are binary and that the state specific conditional response probabilities are time-homogeneous. We set $\theta_{j|1}$ to .75, .8 and .85, $\theta_{j|S}$ to 1-.75, 1-.8, and 1-85, and for $S = 3$, $\theta_{j|2}$ to .58, .65, and .7 which yields the structure shown in Table 1. For $S = 4$, we used the same setting of conditional response probabilities as for $S = 3$, but now defined the conditional response probabilities of the remaining state as high ($=\theta_{j|1}$) for half of the response variables and low ($=\theta_{j|S}$) for the other half.

The design factors and population characteristics were fully crossed resulting in 3 (sample size) $\times$ 2 (number of time points) $\times$2 (number of response variables)$\times$ 2 (number of states) $\times$3 (initial state proportions) $\times$ 3 (transition probability matrices) $\times$ 3 (state-response variables association levels) = 572 simulation conditions. For each simulation condition, a large data set (of 10000 observations) was generated according to the H_1 model and the H_0 parameters were estimated using this data set. Next, for each simulation condition, $K = 1000$ samples were generated according to the H_0 parameters and the CV was computed, assuming $\alpha = .05$. Given a specified sample size, number of time points, and the parameter values under the alternative hypothesis, the power was then computed based on $M = 1000$ samples generated according to the H_1 model as discussed in section 3. To minimize the problem of local maxima, we use multiple random start sets for parameter estimation, in combination with specifying the true population parameter value as the starting value.

4.2 Results

The results obtained from the numerical study for power computation by the short-cut and PBP methods are shown in Tables 2, 3, 4, and 5. As can be seen from these tables, the power values of the two methods are in general comparable. Although the power values obtained by the short-cut method seem to be slightly larger, overall differences do not lead to different conclusions regarding the hypotheses about the number of states. The most important added value of the short-cut method is however that it is $\frac{M}{2}$ times faster than the PBP method, where M refers to the number of Monte Carlo and bootstrap samples for the short-cut and the PBP methods, respectively.

If we now turn to the power values for various combinations of data and population characteristics, we see in Table 2 that the power of the BLR test increases with sample size and the number of time points. Comparison of the effect of sample size and the number of time points show that holding the other factors constant, increasing the number of time points has a larger impact on the power than increasing the sample size. Also, keeping the other design factors constant, the power of the BLR test in general increases with stronger measurement conditions (i.e., weak to moderate to strong state-response variable associations) and with more stable state memberships (smaller transition probabilities).

While in Table 2 we reported the results for equal initial state proportions, in Table 3, we report the results for unequal initial state proportions. As can be seen, the BLR power drops when the initial state size is imbalanced. The more imbalanced the initial state sizes the smaller the power. Table 4 shows the effect of the number of indicator variables on the power of the BLR test: power generally increases when the number of indicator variables increases. Comparing the results in Table 2 with those in Table 5, holding the other factors constant, the power of the BLR test to reject $H_0 : S = 2$ in favour of $H_1 : S = 3$ is in general larger than for $H_0 : S = 3$ against $H_1 : S = 4$.

In summary, the results reported in Tables 2, 3, 4, and 5 show that in the weak measurement condition, the power of the BLR test is in general very low, indicating that very large sample sizes may be required to achieve an acceptable power level in these conditions. Although the quality of state-response association plays a dominant role, the power computed for the weak measurement condition improved substantially by increasing the number of response variables or time points. Also, situations in which the state membership is unstable (e.g., $\rho = 0.3$ or larger) need special care, since the power

is low in such situations.

Table 2: Power of the BLR test for $H_0 : S = 2$ versus $H_1 : S = 3$: the case of equal initial state size ($\delta = 1$) and six indicator variables ($P = 6$)

			State-responses associations								
			Weak			Moderate			Strong		
	sample size	Method	Index of state transition			Index of state transition			Index of state transition		
			$\rho = 0.1$	$\rho = 0.15$	$\rho = 0.3$	$\rho = 0.1$	$\rho = 0.15$	$\rho = 0.3$	$\rho = 0.1$	$\rho = 0.15$	$\rho = 0.3$
$T = 3$	300	short-cut	.188	.145	.104	.301	.260	.176	.568	.494	.339
		PBP	.180	.148	.116	.286	.250	.131	.550	.496	.320
	500	short-cut	.398	.301	.178	.581	.534	.294	.869	.809	.631
		PBP	.394	.280	.150	.558	.493	.296	.858	.804	.610
	700	short-cut	.642	.439	.238	.842	.704	.405	.978	.957	.796
		PBP	.592	.442	.224	.802	.691	.387	.968	.960	.800
$T = 5$	300	short-cut	.698	.559	.228	.849	.727	.394	.972	.927	.687
		PBP	.683	.516	.212	.846	.730	.380	.966	.947	.700
	500	short-cut	.955	.868	.416	.990	.959	.726	1	.999	.942
		PBP	.958	.870	.422	.986	.960	.650	1	.993	.952
	700	short-cut	.995	.970	.654	1	.999	.887	1	1	.992
		PBP	.998	.969	.640	1	.997	.892	1	1	.988

Note. T= number of time points, P= number of response variables, δ=initial state proportion index, ρ=state transition probability index, and PBP= proportion bootstrap p-value rejected.

Table 3: Power of the BLR test for $H_0 : S = 2$ versus $H_1 : S = 3$: the case of imbalanced initial state size ($\delta = 2$ or 3) with six indicator variables ($P = 6$) and three time points ($T = 3$)

			State-responses associations								
			Weak			Moderate			Strong		
	sample size	Method	Index of state transition			Index of state transition			Index of state transition		
			$\rho = 0.1$	$\rho = 0.15$	$\rho = 0.3$	$\rho = 0.1$	$\rho = 0.15$	$\rho = 0.3$	$\rho = 0.1$	$\rho = 0.15$	$\rho = 0.3$
$\delta = 2$	300	short-cut	.147	.130	.080	.247	.197	.135	.346	.308	.249
		PBP	.142	.130	.104	.204	.170	.128	.332	.272	.214
	500	short-cut	.290	.210	.127	.357	.296	.244	.637	.559	.457
		PBP	.274	.208	.132	.336	.280	.252	.604	.544	.410
	700	short-cut	.445	.367	.193	.594	.517	.337	.801	.763	.574
		PBP	.416	.336	.186	.571	.484	.295	.790	.775	.566
$\delta = 3$	300	short-cut	.114	.075	.073	.138	.099	.090	.171	.147	.125
		PBP	.108	.092	.084	.106	.100	.095	.164	.134	.110
	500	short-cut	.146	.112	.104	.196	.173	.131	.307	.281	.220
		PBP	.135	.112	.107	.176	.168	.146	.298	.228	.166
	700	short-cut	.231	.186	.124	.306	.245	.195	.515	.456	.378
		PBP	.193	.176	.124	.289	.246	.196	.482	.374	.244

Note. T= number of time points, P= number of response variables, δ=initial state proportion index, ρ=state transition probability index, and PBP= proportion bootstrap p-value rejected.

Table 4: Power of the BLR test for $H_0 : S = 2$ versus $H_1 : S = 3$: the case of equal initial state size ($\delta = 1$) and three time points ($T = 3$)

			State-responses associations								
			Weak			Moderate			Strong		
	sample size	Method	Index of state transition			Index of state transition			Index of state transition		
			$\rho = 0.1$	$\rho = 0.15$	$\rho = 0.3$	$\rho = 0.1$	$\rho = 0.15$	$\rho = 0.3$	$\rho = 0.1$	$\rho = 0.15$	$\rho = 0.3$
$P = 6$	300	short-cut	.188	.145	.104	.301	.260	.176	.568	.494	.339
		PBP	.180	.148	.116	.286	.250	.151	.550	.496	.320
	500	short-cut	.398	.301	.178	.581	.534	.294	.869	.809	.631
		PBP	.394	.280	.150	.558	.493	.296	.858	.804	.610
	700	short-cut	.642	.439	.238	.842	.704	.405	.978	.957	.796
		PBP	.592	.442	.224	.802	.691	.387	.968	.960	.800
$P = 10$	300	short-cut	.791	.646	.402	.872	.786	.551	.987	.952	.885
		PBP	.766	.618	.394	.848	.746	.540	.970	.948	.848
	500	short-cut	.973	.941	.702	.993	.976	.866	1	1	.993
		PBP	.975	.934	.674	.992	.981	.874	1	1	.992
	700	short-cut	.989	.941	.831	.993	.976	.893	1	1	1
		PBP	.996	.958	.859	1	1	.970	1	1	.998

Note. T= number of time points, P= number of response variables, δ=initial state proportion index, ρ=state transition probability index, and PBP= proportion bootstrap p-value rejected.

Table 5: The power of the BLR test for testing $H_0 : S = 3$ versus $H_1 : S = 4$: the case of equal initial state size and six indicator variables

			State-responses associations								
			Weak			Moderate			Strong		
number of time points	sample size		Index of state transition			Index of state transition			Index of state transition		
			$\rho = 0.1$	$\rho = 0.15$	$\rho = 0.3$	$\rho = 0.1$	$\rho = 0.15$	$\rho = 0.3$	$\rho = 0.1$	$\rho = 0.15$	$\rho = 0.3$
$T = 3$	300	short-cut	.121	.099	.074	.170	.120	.093	.377	.3007	.195
		PBP	.114	.096	.062	.162	.142	.110	.338	.284	.178
	500	short-cut	.199	.158	.122	.272	.230	.171	.643	.539	.341
		PBP	.213	.162	.108	.294	.224	.154	.610	.512	.302
	700	short-cut	.273	.218	.151	.464	.387	.233	.811	.717	.516
		PBP	.258	.222	.126	.474	.354	.238	.794	.718	.498
$T = 5$	300	short-cut	.387	.237	.147	.534	.483	.212	.872	.737	.401
		PBP	.351	.244	.136	.516	.448	.200	.868	.754	.388
	500	short-cut	.738	.551	.214	.882	.802	.361	.994	.962	.706
		PBP	.694	.504	.208	.864	.758	.342	.994	.970	.698
	700	short-cut	.919	.736	.356	.985	.918	.572	1	1	.886
		PBP	.924	.732	.315	.980	.912	.542	1	1	.894

Note. T=number of time points.

Figures 2 and 3 present a power curve (as a function of sample size) for different settings of the parameter values of the 3-state LM population model with equal initial state proportions, 6 response variables, and 3 time points. Figure 2 shows that when the state-response associations are weak, to achieve a power of .8 or larger, we may require a sample of 1000 or more when state membership is stable, and a sample of 2000 or more when state membership is unstable. We can also see from the same figure that when the state-response associations are rather strong, the required sample sizes may drop to less than 500 and 700, respectively for stable and unstable state membership conditions. As can be seen from Figure 3, to achieve a power level of .8 when the state memberships are moderately stable, sample sizes of at least 1200, 850, and 500, may be required in the weak, medium, and strong measurement condition, respectively.

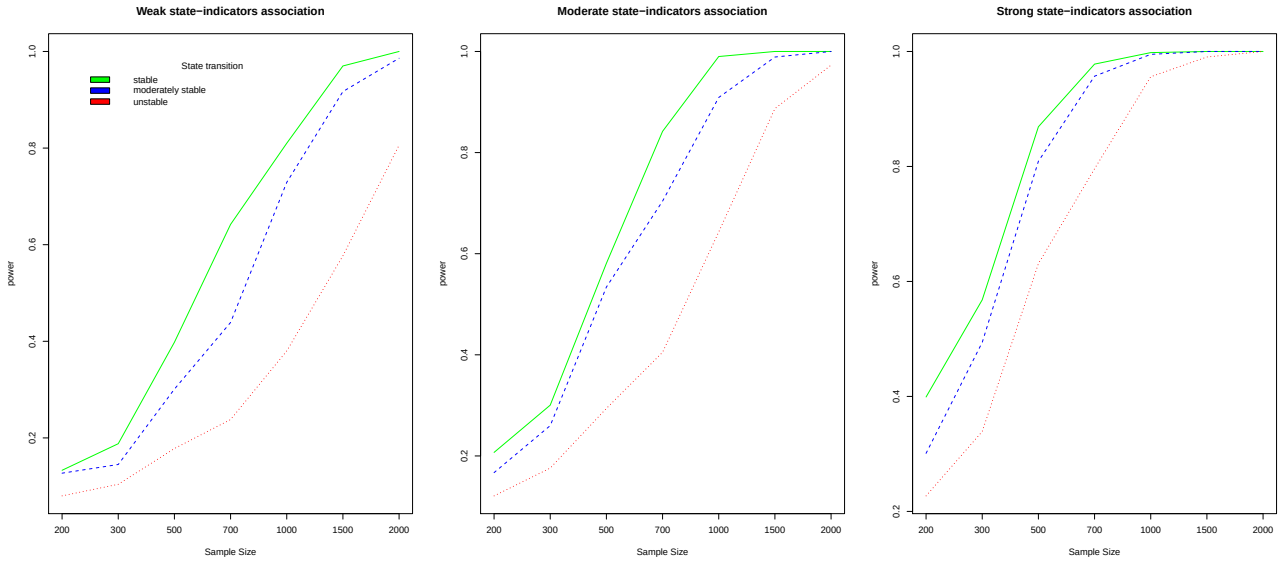


Figure 2: Power by sample size for a 3-state LM population model with varying levels of measurement parameters, equal initial state proportions, 6 response variables, and 3 time points

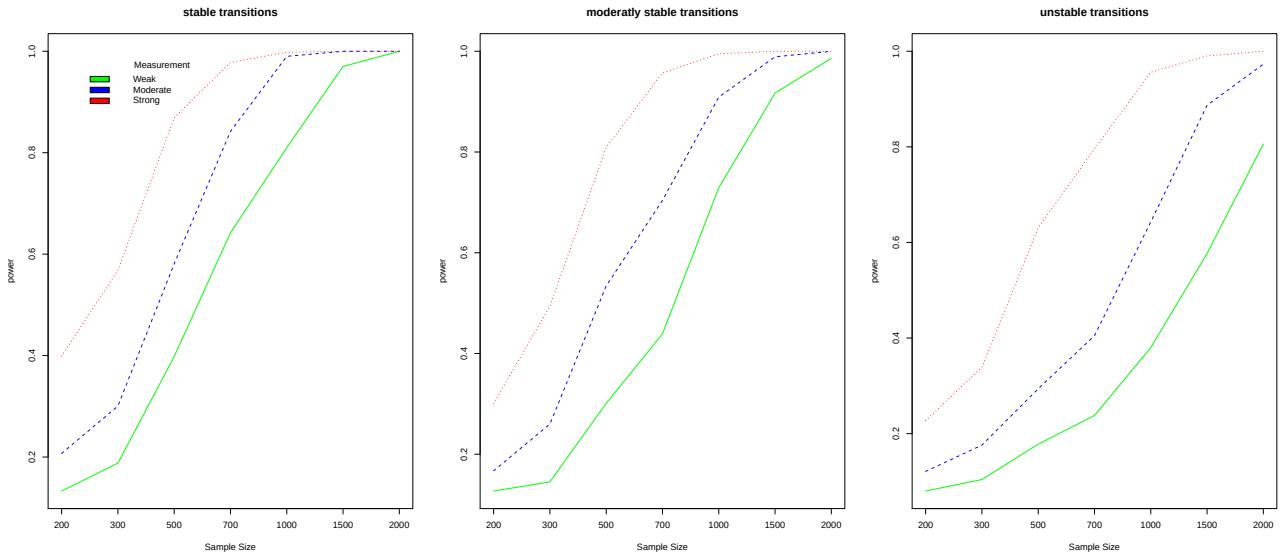


Figure 3: Power by sample size for a 3-state LM population model with varying levels of transition parameters, equal initial state proportions, 6 response variables, and 3 time points

5 Discussion and conclusion

The current study addressed methods of power analysis for the BLR when testing hypotheses on the number of states in LM models. Two alternative methods of power computation were discussed: the proportion of significant bootstrap p-values (PBP) and the short-cut method. Using the PBP method, power is computed by first generating a number of independent data sets under the alternative hypothesis, and then, for each of these data sets, computing the p-value by applying a parametric bootstrap procedure (McLachlan, 1987). The PBP method is computationally very demanding as it requires performing the full bootstrap for each of M samples from the H_1 model. We proposed solving this computational problem using the short-cut method. The short-cut method works very much as a standard power computation, with the difference that instead of relying on the theoretical distributions (a central chi-square under the null hypothesis and a non-central chi-square under the alternative hypothesis), the distributions under H_0 and H_1 are constructed by Monte Carlo simulation.

A numerical study was conducted to (a) illustrate the proposed power

analysis methods and (b) compare the power obtained by the short-cut and the PBP methods. As expected, the power of the BLR test in the LM models increased with sample size. Likewise, power increased with more time points and more response variables. In addition to these design factors, the power of the BLR test was shown to depend on the following population characteristics: the initial state proportions, the state transition probabilities, and the state-response associations. Holding the other design factors constant, power was larger with more balanced initial state proportions, more stable state memberships, and stronger state-response associations. Contrary to this, when initial state proportions are highly imbalanced, state membership is unstable, and the state-response association is weak, the power of the BLR test is low.

For the simulation conditions that we have considered in this study, the sample size required to achieve a power level of .8 or larger ranged from a few hundred to thousands of cases. Also, the required sample size depended on other design factors and population characteristics, which are highly interdependent. In general, the more time points, the more response variables, the more balanced the initial state proportions, the more stable the state memberships, and the stronger the state-response associations, the smaller the sample size needed to achieve a certain power level. Because of mutual dependencies among the LM model parameters, and since the required sample size is also influenced by the number of time points, response variables, and state-indicator variable associations, a sample size of 300 or 500 will often not suffice in LM analysis. Therefore, we strongly suggest applied researchers to perform a power analysis for his/her specific research situation instead of relying on certain rules of thumb about the sample size. The same applies to questions about the minimum number of time points and/or response variables.

Both the short-cut and PBP method discussed in this paper make use of parameter estimates obtained by maximizing the log-likelihood function. In LM models, as in other mixture models, the log-likelihood function can have multiple maxima, meaning that the estimates found do not always correspond to the global maximum of the log-likelihood function. This may have an effect on the computed power (or sample size). In this paper, we dealt with this problem of local maxima by using multiple sets of random starting values for the parameters, in addition to a set of start values corresponding to the known population parameter values.

Limitations to the current numerical experiments need to be acknowl-

edged. Firstly, in the current study, we assumed time homogeneity for both state transition and conditional response probabilities. Future research should assess the power of the BLR test if this assumption is relaxed. Secondly, the conditional response probabilities of the binary response variables were set to equal values, and for simplicity, we considered a specific structure of the transition matrix: $\pi_{s|r} = \frac{\rho^{|s-r|}}{\sum_{h=1}^{S} \rho^{h-r}}$. However, in practice the conditional response probabilities may differ across response variables, the response variables may be nominal with more than two categories, continuous or of mixed type, and the structure of the transition matrix can be completely unconstrained, or, for example, symmetric or triangular (Bartolucci, 2006). Thirdly, this paper focused on power and sample size computation. A further study with more focus on determining the required number of measurement occasions is suggested. Power analysis for the number of time points depends not only the state transition probabilities, but also on the time scale and on whether the dynamics of the system is stationary or not. Fourthly, in our study, we illustrated the proposed power computation methods considering tests for 3-state against 2-state LM models and 4-state against 3-state LM models. In practice, one may encounter tests for larger numbers of states.

It can be concluded that more intensive simulations that address these different scenarios concerning the H_1 population model may be needed to establish more knowledge and guidelines about the power and sample size requirements of the BLR test for the number of states in LM models. What is clear is one should not rely on certain rules of thumb about the required sample size, number of time points, or number of indicator variables, but instead perform a power analysis tailored to the specific situation of interest. The proposed shortcut method makes this computationally feasible.

References

Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society. Series A (General)*, 144(4):419–461.

Bacci, S., Pandolfi, S., and Pennoni, F. (2014). A comparison of some criteria for states selection in the latent markov model for longitudinal data. *Advances in Data Analysis and Classification*, 8(2):125–145.

Bartolucci, F. (2006). Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):155–178.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2010). An overview of latent Markov models for longitudinal categorical data. *arXiv preprint arXiv:1003.2804*.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2013). *Latent Markov models for longitudinal data*. Boca Raton: Chapman and Hall/CRC press.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171.

Brown, B. W., Lovato, J., and Russell, K. (1999). Asymptotic power calculations: description, examples, computer code. *Statistics in Medicine*, 18(22):3137–3151.

Collins, L. M. and Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. New Jersey: John Wiley & Sons.

Collins, L. M. and Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, 27(1):131–157.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 39(1):1–38.

Dias, J. (2006). Latent class analysis and model selection. In *M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nurnberger, & W. Gaul (eds.), From Data and Information Analysis to Knowledge Engineering*, pages 95–102. Berlin: Springer-Verlag.

Feng, Z. D. and McCulloch, C. E. (1996). Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 58(3):609–617.

Gudicha, D. W., Schmittmann, V. D., and Vermunt, J. K. (2015). Power computation for likelihood ratio tests for the transition parameters in latent Markov models. *Structural Equation Modeling: A Multidisciplinary Journal, DOI: 10.1080/10705511.2015.1014040*.

Marsh, H. W., Hau, K.-T., Balla, J. R., and Grayson, D. (1998). Is more ever too much? the number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, 33(2):181–220.

McLachlan, G. (1987). On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Applied Statistics*, 36(3):318–324.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*. New Jersey: John Wiley & Sons.

Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(4):535–569.

O'Brien, R. G. (1986). Using the SAS system to perform power analyses for log-linear models. *Proceedings of the Eleventh Annual SAS Users Group Conference, Cary, NC: SAS Institute*, pages 778–784.

Schmittmann, V. D., Dolan, C. V., van der Maas, H. L., and Neale, M. C. (2005). Discrete latent markov models for normally distributed response data. *Multivariate Behavioral Research*, 40(4):461–488.

Self, S. G., Mauritsen, R. H., and Ohara, J. (1992). Power calculations for likelihood ratio tests in generalized linear models. *Biometrics*, 48(1):31–39.

Steiger, J. H., Shapiro, A., and Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika*, 50(3):253–263.

Vermunt, J. K., Langeheine, R., and Bockenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, 24(2):179–207.

Vermunt, J. K. and Magidson, J. (2013). *LG-Syntax user's guide: Manual for latent GOLD 5.0 syntax module*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., Tran, B., and Magidson, J. (2008). Latent class models in longitudinal research. In *S. Menard (eds.), Handbook of Longitudinal Research: Design, Measurement, and Analysis*, pages 373–385. Burlington, MA: Elsevier.

Yang, C. C. (2006). Evaluating latent class analysis models in qualitative phenotype identification. *Computational Statistics & Data Analysis*, 50(4):1090–1104.