On the use of latent class scores in causal models for categorical variables
Bolck, A.; Croon, M.A.; Hagenaars, Jacques

Publication date:
1998

Link to publication

Citation for published version (APA):
On the use of latent class scores in causal models for categorical variables

WORC Paper 98.07.005/7

Annabel Bolck
Marcel Croon
Jacques Hagenaars
On the use of latent class scores in causal models for categorical variables

Annabel Bolck, Marcel Croon and Jacques Hagenaars

July 16, 1998
On the use of latent class scores in causal models for categorical variables

Annabel Bolck, Marcel Croon and Jacques Hagenaars

July 16, 1998
Introduction

In social sciences, as in many other sciences, often causal relations among variables are analysed. Models most used for causal relations between continuous variables are linear structure equation models. When the variables are categorical, a well-known technique to model causal relations is log-linear analysis. In general, a causal model for categorical variables can be estimated as a whole in a modified path approach. The joint distribution of all variables is decomposed in conditional probabilities based on the causal order of the variables. Eventually, the model can be parameterized with log-linear analysis. Also restrictions can be imposed on the log-linear parameters.

If not all variables are observed, log-linear analysis with latent variables can be used, or in general, a modified path model extended with a component in which the relations between the latent variables and their indicators are specified. This added component is called the measurement part of the model in contrast to the other part, which is called the structural part.

Estimating all parameters (of the measurement part and the structural part) of complex causal models with latent variables simultaneously, requires sophisticated optimization procedures. A major problem is the convergence to suboptimal solutions. Furthermore, the validity of the whole model has to be investigated at once. In the case of misspecification it is not easy to detect which part of the model causes the misspecification. Another problem in the analysis of categorical data is the problem of "sparse tables". Models for categorical data are based on contingency tables. These tables contain as many cells as response possibilities. All cell frequencies have to be estimated for the conditional probabilities. In a relatively simple model which concern 7 trichotomous variables already $3^7 = 2187$ cell frequencies have to be estimated. This can cause serious problems for testing the validity of the model (though not so much for estimating the parameters). Adding 1 trichotomous latent variable that has 6 trichotomous indicators, the number of cells grow to $3^{14} = 4728969$. A gigantic table as this causes many problems, not only with regard to testing the model, but also with regard to estimating the parameters and standard errors.

A possible solution is to apply an approach in two steps, rather than in one step, using two smaller observed tables. First, the measurement model is estimated, followed by the prediction\(^1\) of the individual latent scores.

\(^1\)The term *prediction* is used rather than *estimation* to make a distinction between the
Second, the structural part is estimated in which the individual latent scores are used as if they are observed. In estimating these separate models, more variables can be handled. Only two tables of $3^7$ cells, that is in total 4374 cells, are needed. The techniques used for the separate models are also less prone to local optima, and the validity of the measurement and structural submodel can be investigated separately.

However, nowadays one-step procedures are regarded statistically preferable. Even if measurement models are already estimated in an explorative analyses, it is preferred to estimate all parameters of the causal model simultaneously in one step. This is because of a serious drawback of the two-step procedure, which concerns the relation between the estimated latent scores and variables of the structural submodels. This relation in the two-step procedure is estimated with bias. In this paper, attention is paid to this problem. It is investigated when and why the bias occurs. Moreover, a correction formula is presented to correct for the bias. With this correction the major disadvantage of two-step procedures disappears, which makes it an attractive alternative to the one-step procedure. The approaches are now comparable.

The structure of this paper is as follows: First, a measurement model for categorical data, the latent class model, will be described. Followed, by a description of how individuals can be assigned to latent classes. These together form the measurement part of the two-step procedure. Second, the one-step procedure is compared with the (second step of the) two-step procedure. It is shown that in the second step of the two-step procedure, the strength of the relation between latent and structural variables is in general underestimated. Attention is paid to the question which assignment is best, with respect to the underestimation. After that, a correction formula is provided to correct for the bias in the relation between latent and structural variables in the second step of the two-step approach. Some generalizations are presented of the assumptions made before. Finally, a simulations and a real data example are presented to compare the one-step procedure and the corrected two-step procedure. Pros and cons of both methods are investigated. Some remarks are made about standard errors.
Latent Class Analysis

Latent Class model

Latent Class Analysis (LCA) can be used to estimate a measurement model with categorical data. The classical parameterization of the latent class model in terms of probabilities was introduced by Lazarsfeld (1950b; 1950a) and also used by Goodman (1974a; 1974b). Let $\theta$ be a vector of categorical latent variables with $T$ exhaustive and mutually exclusive latent classes, and $y$ a vector of $p$ categorical (observed) indicator variables, with in total $Y$ exhaustive and mutually independent classes. The realization of $y$ is denoted by $y$, and the 'realization' of $\theta$ by $t$. Let the separate indicator variables be denoted by $A, B, C, \ldots$ with realizations $a, b, c, \ldots$.

\[
p(y) = \sum_{t=1}^{T} p(y = y, \theta = t),
\]

and

\[
p(y, \theta) = p(\theta)p(y|\theta),
\]

where $p(y, \theta)$ is the probability of having response pattern $y$ and latent score $\theta$. Local independence is assumed. Thus the indicators are assumed mutually independent given a particular score on the latent variables. This means, for instance in the case of 1 latent variable and 2 indicators, that Equation 2, with a slightly different notation, can be written as

\[
p(AB\theta) = p(\theta)p(A|\theta)p(B|\theta).
\]

Besides the classical parameterization often a log-linear parameterization of the latent class model is used (Haberman, 1979). Also, further restrictions can be imposed on the parameters. These can be fixed-value or equality restrictions (Goodman, 1974a; Goodman, 1974b; Mooijaart & van der Heijden, 1992), but also inequality restrictions (Croon, 1990). Inequality restrictions can be used for ordinal variables. Categorical variables need not to be nominal. They can be ordinal or discrete interval. In this context, linear-by-linear and column and/or row effects can be brought into a log-linear model (Heinen, 1993; Rost, 1985). However, all these models can in general be represented by the above equations. This is what will be done in the remainder of this paper.

Before the probabilities are estimated, some identifying restrictions are needed. In this paper it is assumed that all identification problems are
solved. Probabilities can be estimated with Maximum likelihood. The best known methods are the Newton-Raphson algorithm and the Expectation-Maximization (EM) algorithm (Dempster, Laird & Rubin, 1977). However, most results presented in this paper, are population results instead of estimations based on samples. These population results naturally have their effect on sample results, which will be shown in a real data example at the end.

Latent Class Assignment

Once the Latent Class model is estimated, unknown individual latent class scores can be predicted. Or stated otherwise, each individual can be assigned to a latent class. Consequently, in this paper the term 'prediction' is used when the latent scores are determined. This in contrast with the term 'estimation' which is used for the determination of model parameters (based on sample data). Furthermore, \( \theta \) is used as the symbol for the latent variable, and \( \tilde{\theta} \) as the symbol for the predicted latent variable. These two, of course have the same number of classes. The 'realization' of the unknown \( \theta \) is denoted by \( t \), and the realization of \( \tilde{\theta} \) by \( s \).

Various assignment rules exist. These can be deterministic, in the sense that individuals with the same response pattern all are assigned to the same latent class, or probabilistic, in the sense that individuals with a particular response have a probability to be assigned to a certain latent class. Like in imputation methods for missing data, drawings are done from the probabilities to determine the class to which an individual will be assigned. In general a Bayesian point of view is taken. All assignment rules are based on the posterior probability \( p(\theta | y) \) that an individual is in latent class \( t \), given (s)he has response \( y \). According to Bayes Theorem

\[
p(\theta | y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y | \theta)p(\theta)}{\sum \, p(y | \theta)p(\theta)}.
\]

Well known assignment methods (Goodman, 1974a; Goodman, 1974b; Lazarsfeld, 1950b; Lazarsfeld, 1950a; Hagenaaars, 1990; Clogg, 1995; Bolck, Croon & Hagenaars, 1997; Bock, 1983) are

- **Modal class assignment or modal a posteriori assignment (MAP):** In this deterministic assignment rule, each individual with response \( y \) on the indicators \( y \) will be assigned to that class \( s \) for wich \( p(\theta = s | y) \) is largest. Thus

\[
p(\tilde{\theta} = s | y) = \begin{cases} 
1 & \text{if } p(\theta = s | y) > p(\theta = s' | y) \forall \ s' \neq s \\
0 & \text{otherwise}
\end{cases}
\]

\[ (4) \]
The total probability of misclassification in modal assignment

\[ E = \sum_y p(y)[1 - p(\theta = s|y)] \]

is minimized.

- **Expected a posteriori assignment (EAP).** This deterministic rule can only be used when the variables are at least ordered. In that case, each individual with response \( y \) is assigned to the mean latent class \( E(\theta|y) \). Thus

\[
p(\tilde{\theta} = s|y) = \begin{cases} 
1 & \text{if } E(\theta|y) = \sum_t \theta \cdot p(\theta|y) = s \\
0 & \text{otherwise}
\end{cases}
\]

(5)

In this assignment the mean-squared error of \( E(\theta|y) \) is minimized.

- **Median a posteriori assignment.** This deterministic rule can only be used when the variables are at least ordered. In that case, each individual with response \( y \) will be assigned to that class \( s \) for which the probability to fall in a lower class given \( y \) equals the probability to fall in a higher class given \( y \) as much as possible.

- **Random assignment:** All individuals with response \( y \) on the indicators \( y \) will be assigned to latent class \( s \) with probability \( p(\theta = s|y = y) \). This means that

\[
p(\tilde{\theta} = s|y) = p(\theta = s|y)
\]

(6)

In this case the total probability of misclassification is:

\[
\sum_y \sum_s p(\theta = s|y)[1 - p(\theta = s|y)]
\]

- **Proportional assignment:** All individuals are proportionally distributed over the latent classes according to the conditional probabilities \( p(\tilde{\theta}|y) \). Thus, \( p(\tilde{\theta} = s|y = y) \times 100\% \) of all individuals with response \( y \) on the indicators \( y \) will be assigned to latent class \( s \). The proportional assignment of an individual to a latent class does not only depend on the response pattern \( y \) of this individual (as in random assignment) but also on the latent classes to which other individuals are assigned. The difference between proportional assignment and random assignment can be seen as the difference between drawing with replacement (random) and drawing without replacement (proportional).
- **Partial assignment**: In this type of assignment, individuals are not assigned to one latent class, but with certain percentages to all classes. These percentages of assignment for an individual with response $y$ equal the probabilities $p(\theta|y = y)$. The total error is the same as in proportional assignment, unless it is assumed that in reality an individual also belongs with certain percentages to latent classes.

The unknown latent class of an individual is predicted with an assignment method, but it is often not exactly the same as the true latent class (Hagenaars, 1990). This is because of a problem which resembles the problem of factor indeterminacy (Steiger & Schonemann, 1978) in Factor Analysis. Even if the observed scores and the exact values of the model parameters are known, this is in general not enough to determine the true latent score for each individual. Unless the probability of misclassification is zero, several sets of latent scores are all in agreement with the model parameters and observed scores. Furthermore, if the probability of misclassification is not equal to zero, the predicted latent scores will not exactly be equivalent to one of the admissible sets of "true" latent scores. Only, if one or more indicators perfectly represents the true latent variable $\theta$, with response probabilities equal to one or zero, and the probability of misspecification equals zero, there is only one admissible set for $\theta$, and $\hat{\theta}$ equals $\theta$.

This all has effects for the relation with external variables in causal models. Because, $\hat{\theta}$ does in general not equal $\theta$ probably the relation with external variables is different for $\hat{\theta}$ and $\theta$. For factor analysis this problem is already investigated (Croon & Boiic, 1997). This paper, will concentrate on methods for categorical variables.

**Comparing one-step and two step procedures of a simple causal model**

A simple causal model is considered for the comparison of one-step and two-step procedures. Later, generalizations to more complex models are made. The model that first will be investigated can be represented by the following diagram:

$$
x \rightarrow \theta \rightarrow y$$
in which $y$ is a vector of categorical indicators, $\theta$ is a vector of categorical latent variables, and $x$ is a vector of categorical external variables. External variables are manifest variables other than indicators. The realization of $x$ is denoted by $x$ and the total number of classes by $X$. The arrows represent a direct effect as in directed graphs. It is assumed that the external variables $x$ do not have a direct effect on the indicators $y$. That is why no arrow points from $x$ to $y$. Note that $x$, $\theta$, and $y$ are vectors, and thus can be multidimensional. In other words, they each can represent more than 1 variable.

The above presented model can be estimated in one step with for instance log-linear analysis. The parameters representing the relation between the latent variables $\theta$ and indicators $y$ (the measurement model), as well as the parameters representing the relation between the external variables $x$ and the latent variables $\theta$ (the structural model) are estimated simultaneously. The relation between $x$ and $\theta$ in general can be represented by the joint or conditional probability of $\theta$ given $x$. The strength of the relation can be represented by a summarizing measure such as the log odds ratio, which is based on the joint or conditional probability. Thus, in general, the model can be written in terms of conditional probabilities by

$$p(x, \theta, y) = p(x)p(\theta|x)p(y|\theta).$$

Note that $p(y|\theta, x) = p(y|\theta)$, because $x$ has no direct effect on $y$. The individual latent scores are not estimated. Thus the probabilities containing $\theta$ are estimated with techniques such as maximum likelihood for latent variables (e.g., iterative procedures such as EM). The model can be parametrized in different ways, but this is not of interest in the context of this paper.

Alternatively, the model can be estimated in two steps. In the first step the measurement model is estimated.

$$\theta \longrightarrow y$$

This can be performed with LCA as described in the foregoing section. Once the model parameters are estimated, individuals are assigned to latent classes according to one of the assignments mentioned. In the second step, the structural model is estimated. The predicted latent scores $\hat{\theta}$ are used as if they are observed in a simple causal model, with no latent variables.
The relation between $x$ and $\theta$ can be represented by the joint or conditional probability of $\theta$ given $x$, or by measures such as the log odds ratio, which are based on the joint or conditional probability. The probability $p(\theta|x)$, can directly be estimated. It is important to note that in the two-step procedure, in contrast to the one-step procedure, the parameters representing the relation between $x$ and $\theta$ are estimated, instead of the parameters representing the desired relation between $x$ and $\theta$.

The predicted latent scores $\tilde{\theta}$ only depend directly on $y$, just as $y$ only depends directly on $\theta$. A graphical representation of the total causal ordering is

\[ x \rightarrow \theta \rightarrow y \rightarrow \tilde{\theta} \]

In terms of conditional probabilities this can be written as

\[ p(x, \theta = t, y, \tilde{\theta} = s) = p(x)p(\theta = t|x)p(y|\theta = t)p(\tilde{\theta} = s|y). \quad (7) \]

Again $p(y|\theta, x) = p(y|\theta)$. Furthermore, it is assumed that $\tilde{\theta}$ is not directly determined by $x$, and from the assignment rules it can be seen that $\tilde{\theta}$ is not directly determined by $\theta$. Thus $p(\tilde{\theta}|x, \theta, y) = p(\tilde{\theta}|y)$. The joint probability for $x$ and $\theta$ is acquired by collapsing over $y$ and $\theta$ of the joint probability of $x$, $\theta$, $y$, and $\tilde{\theta}$ as presented in Equation 7:

\[ p(\tilde{\theta}|x) = \sum_t \sum_y p(\theta|x)p(y|\theta)p(\tilde{\theta}|y) = \sum_t \sum_y p(\theta|x)p(\tilde{\theta}|\theta). \quad (8) \]

From Equation 8 it can be seen that in a two-step procedure the joint (or conditional) probability for $x$ and $\tilde{\theta}$, which is estimated, in general does not equal the joint (or conditional) probability for $x$ and $\theta$, which is desired, and estimated in a one-step procedure. A transition probability $p(\tilde{\theta} = s|x)$ is needed to come from $p(\theta = t|x)$ to $p(\tilde{\theta} = s|x)$.

In the next section it will be shown in what special cases $p(\tilde{\theta}|x) = p(\theta|x)$.

Furthermore, a mathematical proof is provided that in all other cases not only $p(\tilde{\theta}|x) \neq p(\theta|x)$, but also that, in absolute terms, the log odds ratio (as a measure for the strength of a relation between two variables) of $x$ and $\theta$ always underestimates that of $x$ and $\theta$. 
The relation between latent and external variables in a 2-step procedure

True and predicted log odds ratios

Let us denote,
\[ p(e|x) = a_{xt}, \quad p(y|\theta) = b_{ty}, \quad p(\bar{\theta}|y) = c_{ys}, \quad p(\bar{\theta}|\theta) = d_{ts} = b_{ty}c_{ys}, \quad p(\bar{\theta}|x) = e_{xs}. \]

Then
\[ p(\theta, y, \bar{\theta}|x) = a_{xt}b_{ty}c_{ys} \]
and from equation 8:
\[ p(\bar{\theta}|x) = \sum_{t,y} a_{xt}b_{ty}c_{ys} = \sum_{t} a_{xt} \sum_{y} b_{ty}c_{ys} = \sum_{t} a_{xt}d_{ts} = e_{xs}. \] (9)

The last part can be written in matrix notation as

\[ E = AD, \] (10)

with \( \{e_{xs}\} \) the elements of the \((X \times S)\) matrix \(E\), \(\{a_{xt}\}\) the elements of the \((X \times T)\) matrix \(A\), and \(\{d_{ts}\}\) the elements of the \((S \times T)\) transition matrix \(D\). From Equation 10 it can be seen that \(p(\bar{\theta}|x) = p(\theta|x)\) for all values of \(x\) and \(\theta\), only if \(E\) equals \(A\), or, stated otherwise, when \(D = I\). As noted before, the number of classes \(S\) equals the number of classes \(T\). This is true when the indicators represent the latent classes perfectly, and each latent class is assigned to with probability 1 given at least one response pattern\(^2\). In this case the total error is zero. In general \(p(\bar{\theta}|x) \neq p(\theta|x)\).

Let us now examine in what way measures for the relation between \(x\) and \(\bar{\theta}\) differ from measures for the relation between \(x\) and \(\theta\). Conditional probabilities represent the relation between variables at various values of the variables. A summarizing measure for the strength of the relation in the case of categorical variables is the log odds ratio. The true odds ratio, representing the strength of the relation between the classes \(x_1\) and \(x_2\) of \(x\), and the classes \(t_1\) and \(t_2\) of \(\theta\) is

\[ \alpha_{x_1x_2t_1t_2} = \frac{a_{x_1t_1}a_{x_2t_2}}{a_{x_1t_2}a_{x_2t_1}}. \] (11)

\(^2\)If \(b_{ty} \geq 0 \Rightarrow (b_{ty} = 0) \land (c_{yt} = 1) \land (c_{yt'} = 0)\); if \(b_{ty} = 0 \Rightarrow (b_{ty} \neq 0) \land (c_{yt} = 0) \land (c_{yt'} = 1)\); \forall t b_{ty} > 0 for at least one \(y\).
While, the predicted odds ratio representing the relation between the (classes $x_1$ and $x_2$ of the) external variable $x$ and the (classes $t_1$ and $t_2$ of the) predicted latent variable $\theta$ is

$$\hat{\alpha}_{x_1 x_2 t_1 t_2} = \frac{e_{x_1 t_1} e_{x_2 t_2}}{e_{x_1 s_1} e_{x_2 s_2}} \quad (12)$$

$$= \frac{\sum t_1 a_{x_1 t_1} d_{t_1 s_1} \sum t_2 a_{x_2 t_2} d_{t_2 s_2}}{\sum t_1 a_{x_1 t_1} d_{t_1 s_1} + \sum t_2 a_{x_2 t_2} d_{t_2 s_2}}$$

$$= \frac{\sum t_1 a_{x_1 t_1} d_{t_1 s_1} \sum t_2 a_{x_2 t_2} d_{t_2 s_2} \sum t_1 a_{x_2 t_1} d_{t_1 s_2}}{\sum t_1 a_{x_1 t_1} d_{t_1 s_2} \sum a_{x_2 t_1} d_{t_2 s_2} \sum a_{x_1 t_1} d_{t_1 s_2} + \sum a_{x_2 t_2} d_{t_2 s_2} \sum a_{x_1 t_1} d_{t_1 s_2}} \quad (13)$$

In Theorem 1 below, it is stated that in the case of a dichotomous latent variable $\theta$ and arbitrary $x$ and $y$, the strength of the relation between $x$ and $\theta$, is always underestimated, in the two-step procedure. The predicted log odds ratio is, in absolute terms, always smaller than the true log odds ratio.

**Theorem 1** When $\theta$ is dichotomous, then: $|\log(\hat{\alpha}_{x_1 x_2})| \leq |\log(\alpha_{x_1 x_2})|$, or stated otherwise $\hat{\alpha}_{x_1 x_2} \leq \alpha_{x_1 x_2}$ for $\alpha_{x_1 x_2} \geq 1$ and $\hat{\alpha}_{x_1 x_2} \geq \alpha_{x_1 x_2}$ for $\alpha_{x_1 x_2} \leq 1$.

**Proof.** The proof is based on another theorem (see also Appendix A):

**Theorem 2** Let $u_i$, $v_i$ and $w_i$ be a series of numbers for which $u_i \geq 0, v_i \geq 0, w_i > 0$ and at least one $v_i > 0$ Then

$$\min_i \frac{u_i}{w_i} \leq \sum_i \frac{u_i v_i}{v_i w_i} \leq \max_i \frac{u_i}{w_i}. $$

In formula 13, $a_{x_1 t_1} d_{t_1 s_1} a_{x_2 t_2} d_{t_2 s_2}$ can be considered as $u_i v_i$, and $a_{x_1 t_2} d_{t_2 s_2} a_{x_2 t_1} d_{t_1 s_1}$ as $v_i w_i$. They are summed over $t_1 = 1$ and $t_2 = 2$, which is equivalent to $i = 1, \ldots, 4$. It can be seen that $a_{x_1 t_1} d_{t_1 s_1} a_{x_2 t_2} d_{t_2 s_2} = a_{x_1 t_1} a_{x_2 t_2} a_{x_1 t_2} a_{x_2 t_1}$ (or $\frac{u_i v_i}{v_i w_i} = \frac{u_i}{w_i}$).

The 4 possible values of $\frac{u_i}{w_i}$ are

$$\frac{a_{x_1 t_1} a_{x_2 t_2}}{a_{x_1 t_1} a_{x_2 t_1}} = 1, \text{ when } t_1 = 1; t_2 = 1$$

$$\frac{a_{x_1 t_1} a_{x_2 t_2}}{a_{x_1 t_2} a_{x_2 t_1}} = \alpha_{x_1 x_2}, \text{ when } t_1 = 1; t_2 = 2$$

$$\frac{a_{x_2 t_1} a_{x_2 t_2}}{a_{x_1 t_1} a_{x_2 t_1}} = \frac{1}{\alpha_{x_1 x_2}}, \text{ when } t_1 = 2; t_2 = 1$$

$$\frac{a_{x_2 t_1} a_{x_2 t_2}}{a_{x_1 t_2} a_{x_2 t_1}} = \frac{1}{\alpha_{x_1 x_2}}, \text{ when } t_1 = 2; t_2 = 1$$
\[
\frac{a_{x1}a_{x2}}{a_{x1}a_{x2}} = 1, \text{ when } t_1 = 2; t_2 = 2
\]

If it is supposed that at least one \((d_{t_1s}, d_{t_2s}) \neq 0\)

i) If \(\alpha_{x1}x_2 \geq 1 \Rightarrow 1/\alpha_{x1}x_2 \leq 1\), then according to Theorem 1: \(1/\alpha_{x1}x_2 \leq \alpha_{x1}x_2\).

ii) If \(\alpha_{x1}x_2 \leq 1 \Rightarrow 1/\alpha_{x1}x_2 \geq 1\), then according to Theorem 1: \(\alpha_{x1}x_2 \leq 1/\alpha_{x1}x_2\)

A generalization can be made when \(\theta\) has more than 2 categories. Suppose \(\theta\) is trichotomous, then \(\bar{\theta}\) is also trichotomous. The \(t_1\) and \(t_2\) can take, for instance, the values 0, 1 and 2. Then

\[
\frac{e_{x1s1}e_{x2s2}}{e_{x1s2}e_{x2s1}} = \frac{a_{x10d_0s1}a_{x20d_0s2} + \ldots + a_{x12d_2s1}a_{x22d_2s2}}{a_{x10d_0s2}a_{x20d_0s1} + \ldots + a_{x12d_2s2}a_{x22d_2s1}} = \frac{a_{x10}a_{x20} + \ldots + a_{x12}a_{x22}}{a_{x10}a_{x20} + \ldots + a_{x12}a_{x22}},
\]

which lies, according to theorem 2, between the minimum and maximum of \(\left\{\frac{a_{x10}a_{x20}}{a_{x12}a_{x22}} \ldots \frac{a_{x12}a_{x22}}{a_{x10}a_{x20}}\right\}\). In other words, in the case of more than two latent categories it is possible that not all predicted log odds ratios, describing the strength of the relation between latent and external variables, are underestimated. However, no predictions are larger than the largest true log odds ratio!

**Which assignment is best?**

In the previous sections is proved that, whatever the type of latent class assignments is, the relation with external variables is usually underestimated. Apart from this it is interesting to know what the performance of the different types of assignments is. In the section about latent class assignment it was already stated that the total probability of misclassification is lowest for modal class assignment. There is however not a direct relation between the total probability of misclassification and the difference between \(p(\theta|x)\) and \(p(\bar{\theta}|x)\). In theory, the total probability can be very high while on the other hand \(p(\bar{\theta}|x)\) exactly equals \(p(\theta|x)\). However, in the dichotomous case till now it was always found that modal assignment resulted in log odds

\[3(d_{t_1s1}d_{t_2s2}) = 0 \forall t, s, \text{ if the probability of assigning to } \bar{\theta} = t \text{ given a response } y \text{ is 0 when in all cases the probability of having that response } y \text{ given a latent class } \theta \text{ does not equal 0.} \]
ratios which resemble the true log odds ratios more than the log odds ratios estimated after proportional or random assignment. The proof is a subject for further research.

A correction procedure

Above, it is shown that in estimating the relation between latent and external variables with a two step procedure, not the relation between the external variables $x$ and the true latent variables $\theta$ is estimated, but the relation between $x$ and the predicted latent variables $\hat{\theta}$. The relation between $\theta$ and $x$ can be represented by $p(\theta|x)$, and be written as elements of the matrix $A$. The relation between $\hat{\theta}$ and $x$ can be represented by $p(\hat{\theta}|x)$, and be written as elements of the matrix $E$. It was also shown that $E = AD$, and that usually $D$ does not equal $I$, or stated otherwise, that $E$ does not equal $A$. However, all elements of $D$ are estimated. Thus, if $D$ is invertable, the true conditional probabilities of the latent variables given the external variables can be calculated by

$$A = ED^{-1}. \quad (14)$$

This formula states that, if individuals are assigned to classes with whatever assignment rule that results in a non-singular $D$, the relation between external and latent variables may be underestimated, but can be corrected afterwards such that a good estimation of the relation is obtained. This is true for dichotomous and polytomous variables, because the matrices $E, A$ and $D$ can be as large as desired.

$D$ can become singular when for instance no individuals are assigned to a certain class. Then $D$ has a column with only zeros. In a situation with a large number of classes, this is a possibility. Restrictions, a general inverse, or just an assignment rule that does not lead to a singular $D$ can be used to overcome the problem of singularity.

Note that it does not matter, for the use of the correction procedure, what assignment rule is used, as long as the assignments are dependent of the response pattern (and $D$ is not singular). This means that for instance a simple rule as

$$\hat{\theta} = \begin{cases} 1 & \text{if } y \in Y_1 \\ 2 & \text{if } y \in Y_2, \end{cases} \quad (15)$$

with $Y_1$ some subset of all responses $y$, and $Y_2$ the complement of $Y_1$, can be used.
As an illustration, \( A = ED^{-1} \) is written in the dichotomous case as
\[
\begin{pmatrix}
a_{11} & a_{12} \\
a_{21} & a_{22}
\end{pmatrix} = \frac{1}{|D|} \begin{pmatrix}
e_{11} & e_{12} \\
e_{21} & e_{22}
\end{pmatrix} \begin{pmatrix}
d_{22} & -d_{12} \\
-d_{21} & d_{11}
\end{pmatrix}.
\]
The matrix \( D \) is singular when \( d_{11} = d_{21} \) (and thus \( d_{12} = d_{22} \)). This is the case when \( b_{1y} = b_{2y} \) forall \( y \), that is, when the probability of having a response \( y \) is the same for latent class 1 and latent class 2. The indicators do not represent the latent variable (there is zero correlation between \( y \) and \( \theta \)). This is a situation, for which no good assignment rule exists, but which may never occur in practice when the indicators are well chosen.

The dichotomous matrix \( D \) can also be singular when all individuals are assigned to only one latent class. This probability is, however, very low. In the case of polytomous variables it can happen, with certain assignment rules, that no individuals are assigned to one of the latent classes. This always has to be checked. And it always has to be tried to find some assignment which lead to a non-singular \( D \). In general, not much is required. It is usually enough that the assignment is dependent of the response \( y \).

**Generalization**

Till now only a simple model in which the external observed variables \( x \) are exogenous, and the latent variables \( \theta \) are endogenous, is considered. However, all kinds of relations between latent and external (observed, other than the indicators) variables may exist. The correction formula in the two-step procedure can be used in most relations. The only restriction to use the correction formula on a model is that it is assumed that there are no relations between the indicators and the external observed variables. Figure 1 provides a graphical representation of all possible relations between external and latent variables. The symbols \( x, z, \theta_1, \theta_2 \) represent vectors (thus can be multi-dimensional) of exogenous and endogenous external and latent variables respectively, \( y_1 \) and \( y_2 \) represent vectors of indicators, and \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) represent vectors of predicted latent variables. Between the variables no arrows are drawn because the relations can lie in both directions, except the relations between \( \theta, y, \) and \( \theta \). The indicators \( y \) are only directly determined by \( \theta \), and the predictions \( \hat{\theta} \) are only directly determined by \( y \). The numbers refer to the numbers of the relations as they will be quoted below. All possible relations between (external) observed variables and latent variables can, in fact, be reduced to three types:
Relations between external variables (e.g., relation 5) can directly be estimated. There will be no difference between the two-step procedure and the one-step procedure.

Relations between external and latent variables (e.g., 1, 2, 3, and 4). This is the type already described in this paper. Till now, however, the conditional probability, in which the latent variable depends on the external variable, is concerned. In general the joint distribution between a latent variable \( \theta \) and an external variable \( u \) can be considered. In this case, it does not matter whether \( u \) determines \( \theta \), or \( \theta \) determines \( u \).

For the calculation of the log odds ratios conditional as well as joint distributions can be used. In Figure 1, \( u \) stands for \( x \) or \( z \), and \( \theta \) stands for \( \theta_1 \) or \( \theta_2 \). The relation between \( u \) and \( \theta \) can in the one-step procedure be represented by \( p(\theta, u) \), and in the two-step procedure by \( p(\bar{\theta}, u) \). This last probability can be written in terms of \( p(\theta, u) \):

\[
\begin{align*}
e_{us} = p(\theta = s, u) &= \sum_t p(\theta = t, u) \sum_y p(y|\theta = t)p(\bar{\theta} = s|y) \\
&= \sum_t p(\theta = t, u)p(\bar{\theta} = s|\theta = t) = a_{ut}d_{ts},
\end{align*}
\]

or in matrix notation \( E = A.D. \).
• Relations between latent variables (e.g., relation 6) can in a one-step approach be represented by \( p(\theta_1, \theta_2) \), and in a two-step approach by \( p(\theta_1, \theta_2) \). This can be written in terms of \( p(\theta_1, \theta_2) \) as:

\[
p(\hat{\theta}_2 | \theta_1) = \sum_{y_1} \sum_{y_2} p(\theta_2 | \theta_1, y_1) p(\theta_1 | y_1) p(y_2 | \theta_2) p(\hat{\theta}_2 | y_2)
\]

Thus

\[
E = H^T AD
\]

The external variables (x) have no influence on this relation, and therefore are summed out.

In general, it can be stated that if \( E \) is the matrix of the predicted joint or conditional distributions representing the relation between two variables, and \( A \) is the matrix of the desired joint or conditional distributions, then:

\[
E = H^T AD,
\]

with \( H \) a matrix belonging to the independent variable and \( D \) a matrix belonging to the dependent variable. If both variables are observed \( H = D = I \), if one variable is observed only one of \( H \) or \( D \) equals \( I \), and if both variables are latent both \( H \) and \( D \) do not equal \( I \). A correction formula can be computed by

\[
A = (H^T)^{-1} ED^{-1}
\]

Only when there are relations between external variables and the indicators, no correction formula exists. If there are two (or more) groups of latent variables with (partly) overlapping indicators, these two groups can be seen as one large vector of latent variables with one large vector of indicators of which some (partly) represent other latent variables than others.
Examples

The theory presented above concerns population theory. In practice samples are drawn to estimate population values such as the joint and conditional probabilities. Results found for population values, such as the fact that the one-step conditional probability between a latent and external variable, and the two-step conditional probabilities deviate, and that it is possible to correct for this deviation, still can be used but results will be less exact because of estimations and rounding errors. To illustrate this and to make a comparison between the one-step and the corrected two-step procedure, two examples are presented. First a large simulated data set, is presented, in which the problems of the one-step and two-step procedures are illustrated. Second, a (smaller) real data example is presented, of which is known that the models estimated are stable and good. This example is presented to demonstrate the correction of the two-step estimations of the log odds ratio's between latent variables and external variables. It is followed by some remarks about the standard error.

A simulation example

Figure 2, represents the population which is simulated. The latent variable $\theta_1$ consists of 3 latent classes, and latent variable $\theta_2$ of 4 latent classes.
The external variables \( x, y, \) and \( z \) have 6, 4, and 2 categories respectively. There are eight indicators (denoted by \( \text{A to H} \)) for latent variable \( \theta_1 \) with 2, 2, 2, 2, 3, 3, 3, and 6 categories respectively, and eight indicators (denoted by \( \text{I to P} \)) for \( \theta_1 \), with 2, 2, 2, 2, 4, 4, 4, and 6 categories respectively. A sample of 1000 is drawn from the population. This sample is considered not too large in the sense that it will be almost exactly the population and not too small in the sense that this magnitude would be considered desirable if the data were real. Note that though all variables are denoted bold, which is a vector notation, they are only one-dimensional.

The model as presented above is estimated in one and two steps. The estimation is performed in \( \ell \)EM (Vermunt, 1997). Estimating the model with a one-step procedure costs a few minutes\(^4\), and results in every run to other solutions. There exist many local solutions. All solutions show the same trend as the population, but differ with each other, with respect to the boundary solutions. Table 1 provides the log odds ratio for the relation between the external variable \( z \) and the latent variable \( \theta_1 \) for the population, for 2 random runs of the one-step procedure, and a run with the population values as starting values (run 3). For reasons of simplicity, the relation between \( z \) and \( \theta_2 \) is chosen as an illustration, because it has the smallest number of log odds ratios. The simulated conditional probabilities of \( z|\theta_2 \) were

\[
\begin{align*}
p(z = 1|\theta_2 = 1) &= 0.3 \quad p(z = 2|\theta_2 = 1) = 0.7 \\
p(z = 1|\theta_2 = 2) &= 0.5 \quad p(z = 2|\theta_2 = 2) = 0.5 \\
p(z = 1|\theta_2 = 3) &= 0.7 \quad p(z = 2|\theta_2 = 3) = 0.3 \\
p(z = 1|\theta_2 = 4) &= 0.9 \quad p(z = 2|\theta_2 = 4) = 0.1.
\end{align*}
\]

In the following, the runs with the population values as starting values will be used as reference. It is the model most close to the population.

In the two-step procedure, first the latent class model for the the latent variable \( \theta_1 \) and separately that for the latent variable \( \theta_2 \), are estimated, with the population values as starting values. LCA on \( \theta_1 \) always provides the same results. LCA on \( \theta_2 \) provides a few different solutions which result in almost no differences in the modal assignments based on the models. The second step in the two-step procedure always provides the same result if the same assignments are used. If the second step was based on different assignments, different results were obtained, but of a smaller scale than in

\(^4\)When the standard errors are estimated it takes about 2 hours on a standard pentium 100
z/θ2 | population | 1-step (run1) | 1-step (run2) | 1-step (run 3)  
---|---------|---------|---------|---------
ln α_{12} | -0.85 | -0.30 | -0.36 | -0.32 |
ln α_{13} | -1.69 | -1.11 | -1.95 | -1.30 |
ln α_{14} | -3.04 | 2.13 | -2.43 | -2.34 |
ln α_{23} | -0.85 | 0.19 | -1.59 | -0.98 |
ln α_{24} | -2.20 | -1.90 | -2.07 | -2.02 |
ln α_{34} | -1.35 | -2.10 | -0.48 | -1.04 |

Table 1: The log odds ratios of the population values of the relation between z and θ_2 and the estimated log odds ratios in 3 separate log-linear analyses (1-step procedures) on a sample of 1000 out of the population. Of 10 runs, run 1 provided the lowest χ^2 and run 2 the highest. Run 3 provides the results of the log-linear analysis when the population values are used as starting values.

the one-step procedure. In contrast to the one-step procedure there were only a few different end solutions.

Table 2, provides the log odds ratio’s for the relation between the external variable z and the latent variable θ_2 in the population and the log odds ratios estimated in the two-step procedure, before and after correction with the correction procedure described in this paper. From Table 2 (and Table 1) it can be seen that the log odds ratios estimated in the two-step procedure are worse than those estimated in run 3 of the one-step procedure when compared with the population. After correction with the procedure.

z/θ2 | population | 2-step before correction | 2-step after correction  
---|---------|----------------|----------------|
ln α_{11} | -0.85 | -0.23 | -0.31 |
ln α_{12} | -1.69 | -1.42 | -1.40 |
ln α_{13} | -3.04 | -1.91 | -2.20 |
ln α_{23} | -0.85 | -1.18 | -1.09 |
ln α_{24} | -2.20 | -1.67 | -1.89 |
ln α_{34} | -1.35 | -0.49 | -0.80 |

Table 2: The log odds ratios of the population values of the relation between z and θ_2, and the log odds ratios estimated in the two-step procedure (with the population values as starting values) before and after correction.
presented in this paper, the solutions are improved.

In practice, no population information is known. A one-step or a two-step procedure will be performed to estimate the model. Some remarks can be made. Especially in the case of a one-step procedure often various solutions are found when the estimation is repeated. It is not easy to find out which one is best. Furthermore, in this paper is shown that the estimated parameters in the two-step procedure are underestimations, but also that it is possible to correct for this. Before a model is estimated, however, a good model has to be found. Therefore, an interesting question is: How do the one-step and two-step procedure perform with respect to finding the correct model? In practice you cannot check this, but in the above simulation example we exactly know the population, and can thus really check this.

Suppose researchers of the sample of 1000 described above, have some hypotheses about the model for this data which deviate from which we know is the correct model. Three deviating models are considered. Below is described in what aspect(s) these three differ from the population model:

1. No relation is assumed between $y$ and $\theta_2$
2. No relation is assumed between $y$ and $\theta_2$, and it is assumed that $\theta_1$ is directly caused by $x$ and $y$, with $x$ and $y$ independent.
3. The indicators $A, B$, and $H$ are left out of the estimation

In the first two cases it was found that the models did not fit the data very well in the one-step procedure and in the two-step procedure. In the two-step procedure no fit problems were found in the latent class models but only in the log-linear analysis in the second step. In the third case were 3 important indicators for $\theta_1$ were left out, there were no lack of fit problems found in the one-step procedure though the latent class model for $\theta_1$ showed that the 5 indicators left did not fit the LCM at all.

A real data example

As a real data example a not too large example is chosen, of which is known that the model is stable and will not be rejected. This problem is presented only to illustrate the correction procedure described in this paper. No problems such as described in the simulation model above occur. Results from the one-step and two-step procedures are, therefore, stable and can be compared. The estimated log odds ratios of the one-step and two-step procedures will, however, not be exactly equal to the true (but unknown)
Table 3: A real data example: The 1241 German respondents on the political action study. A, B, C, and D are the indicators, and R, I, and P the external variables, for explanation see text.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>R=1</th>
<th>R=1</th>
<th>R=1</th>
<th>R=1</th>
<th>R=2</th>
<th>R=2</th>
<th>R=2</th>
<th>R=2</th>
<th>tot</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>15</td>
<td>5</td>
<td>14</td>
<td>18</td>
<td>7</td>
<td>13</td>
<td>16</td>
<td>94</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>21</td>
<td>7</td>
<td>17</td>
<td>22</td>
<td>13</td>
<td>18</td>
<td>12</td>
<td>128</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>7</td>
<td>5</td>
<td>34</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>8</td>
<td>1</td>
<td>8</td>
<td>6</td>
<td>38</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>9</td>
<td>0</td>
<td>4</td>
<td>16</td>
<td>8</td>
<td>15</td>
<td>10</td>
<td>64</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>15</td>
<td>59</td>
<td>11</td>
<td>32</td>
<td>53</td>
<td>41</td>
<td>33</td>
<td>39</td>
<td>283</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>12</td>
<td>2</td>
<td>20</td>
<td>15</td>
<td>13</td>
<td>19</td>
<td>12</td>
<td>97</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>10</td>
<td>33</td>
<td>6</td>
<td>17</td>
<td>27</td>
<td>22</td>
<td>23</td>
<td>25</td>
<td>163</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>44</td>
<td>14</td>
<td>44</td>
<td>26</td>
<td>39</td>
<td>41</td>
<td>249</td>
<td></td>
</tr>
<tr>
<td>tot</td>
<td>65</td>
<td>219</td>
<td>54</td>
<td>160</td>
<td>218</td>
<td>139</td>
<td>200</td>
<td>186</td>
<td>1241</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

log odds ratios, and also not equal each other, because we are dealing with samples. Though the correction of the predicted probabilities in the population case will exactly provide the true probabilities, this will now also not be the case, because we are dealing with estimated predicted probabilities based on samples.

The real data example was presented before by Hagenaars (1990, pp137-142), in a slightly different form and without correction. The data come from a political action study, an international comparative survey conducted in eight countries (Barnes & Kaase, 1979). It was investigated whether the different perspectives people have on government tasks leads to different political party preferences (P). A distinction is made between a left wing (1) and centered right wing (2) preference. Two background variables, Religiousness (R) and Income characteristics (I) are concerned (for both: 1=low and 2=high). The perspective people have on government tasks is the latent
variable ($\theta$). It has three categories. When people think that the government has an essential task in "ideal" and "material" matters $\theta = 1$, in "material" but not in "ideal" matters $\theta = 2$, and has no essential responsibilities $\theta = 3$. Respondents were asked which of a list of ten issues have to be considered as a responsibility of government. Four issues are selected to serve as indicators (A: guaranteeing equal rights for men and women; B: providing a good education; C: providing a good medical care; D: Providing equal rights for guest workers). The indicators are dichotomous, a person concerns it an essential responsibility of the government (1) or not (2). The data of 1241 respondents in Germany are summarized in Table 3, and the model considered is presented in Figure 3. As can be noted, a slightly different notation for the variables is used.

In Hagenaars (1990) the above model is estimated with log-linear analysis, a one-step. The model \{RI, I\theta\}\{RI\theta, RIP, \theta P, \theta A, \theta B, \theta C, \theta D\} is found to be the best log-linear model. Also a two-step procedure was used. For the measurement part to be identified, it is assumed that people who only think the government has an essential task in material matters find ideal matters as less essential as people who think the government has no essential responsibilities ($p(A|\theta = 2) = p(A|\theta = 3)$ and $p(D|\theta = 2) = p(D|\theta = 3)$), and find material matters as important as people who think both material and ideal matters are important ($p(B|\theta = 2) = p(B|\theta = 1)$ and $p(C|\theta = 2) = p(C|\theta = 1)$).

In this paper the same one-step and two-step models as in (Hagenaars,
Tables 4 and 5 provide the estimated log odds ratios for the relations between $\theta$ and $I$, and $\theta$ and $P$ respectively.

In the two-step procedure first a Latent Class Analysis, with the same restrictions as described above, is performed on the indicators. This is followed by a modal class assignment as well as a random assignment. In both cases the predicted individual latent scores are used, as if they are observed, in a loglinear analysis with Religiousness ($R$), Income ($I$) and Party preference ($P$), in the same structural model as in the one-step procedure. The

1990) are estimated in EEM (Vermunt, 1997). In the one step procedure the parameters of the causal model $\{RI, I\theta\}$ are estimated simultaneously (Pearson $\chi^2 = 100.63$; $L^2 = 104.39$; 106 degrees of freedom; $p = 0.53$). The same log-linear parameters are found as in (Hagenaars, 1990). Tables 4 and 5 provide the estimated log odds ratios for the relations between $\theta$ and $I$, and $\theta$ and $P$ respectively.

In the two-step procedure first a Latent Class Analysis, with the same restrictions as described above, is performed on the indicators. This is followed by a modal class assignment as well as a random assignment. In both cases the predicted individual latent scores are used, as if they are observed, in a loglinear analysis with Religiousness ($R$), Income ($I$) and Party preference ($P$), in the same structural model as in the one-step procedure. The
estimated predicted log odds ratios after modal and random assignment are denoted in Tables 4 and 5.

The estimated probabilities on which the estimated predicted log odds ratios are based are corrected with the correction formula of Equation 14. The resulting corrected estimated predicted log odds ratios are also denoted in Tables 4 and 5.

From Tables 4 and 5 it can easily be seen that the log odds ratios in the two-step procedure are lower than the largest log odds ratio in the one-step procedure. The one-step procedure log odds ratios are not the 'true' log odds ratios, but are estimations of it. The predicted two-step log odds ratios are underestimations of the true log odds ratios. Especially in the relation between \( \theta \) and \( I \) it can be seen that modal assignment performs better than random assignment. After correction the estimated predicted log odds ratios of the two-step procedure are improved a lot when compared with the estimated log odd ratios of the one-step procedure.

**Standard errors**

When samples are used instead of populations, standard errors become important. Some remarks about standard errors will be made in this section. The asymptotic standard error of the log odds ratio between the (classes \( t_1 \) and \( t_2 \) of the) latent variable \( \theta \) and the (classes \( x_1 \) and \( x_2 \) of the) external variable \( x \) can be computed with the delta method by

\[
\sigma \log(p_{x_1x_2t_1t_2}) = \left( \frac{1}{n_{a_{x_1t_1}}} + \frac{1}{n_{a_{x_2t_2}}} + \frac{1}{n_{a_{x_2t_1}}} + \frac{1}{n_{a_{x_2t_2}}} \right),
\]

with \( a_{xt} = p(x = x, \theta = t) \).

In the same way, a prediction of the standard error in a two-step procedure can be computed by filling in the estimated predicted joint probabilities \( p(x, \hat{\theta}) \). As expected, this standard error underestimates the 'true' standard error. A corrected form can be obtained by filling in the corrected estimated predicted joint probabilities. This, however, still underestimates the standard error in the one-step procedure. In a one-step procedure the latent variable is really latent, thus not known or predicted as in the two-step procedure. Therefore, the estimation for \( p(\theta|x) \) in the one-step procedure contains more uncertainty than that for \( p(\theta|x) \). Not only the probability, but also the corresponding "observed" proportion is unknown.

\[\text{Var}(\text{row of } A) = (D^{-1})' \text{Var}(\text{row of } E) D^{-1},\]

which is exactly the same.
When parameters are estimated with a maximum likelihood method the standard error can be estimated based on the information matrix. An estimation of the one-step standard error without estimating the one-step procedure can be provided by filling in the corrected estimated conditional probabilities in the theoretical information matrix of a one-step procedure. However, the (theoretical) information matrix can become very complex.

Approximations of the standard error based on the two-step procedure that come close to the values of the one-step procedure can be provided with Multiple Imputation (Rubin, 1987), a technique developed for missing data. In missing data theory a lot of possibilities exist to replace missing data by imputed values. In Multiple imputation, the missing data are not replaced by one set of imputations but by several \( m \) sets of imputations, resulting in \( m \) completed datasets on each of which complete data set analyses are performed.

In this paper, a single imputation concerns assignment of all individuals to latent classes. Based on this imputation the second step of the two-step procedure is performed, resulting in log odds ratios for the relation between the latent variables, and/or the latent variables and external variables. These log odds ratios can be corrected with the correction procedure and a standard error can be estimated. This whole sequence of procedures is repeated \( m \) times starting each time with a new assignment. In this way not only \( m \) standard errors or variances of the log odds ratios are computed, but also the variance between the \( m \) (sets) of log odds ratios can be computed. This last variance is denoted by \( B \), and the mean of the \( m \) variances of the log odds ratios (the within variance) is denoted by \( U \). The multiple imputation standard error is then computed by

\[
\sqrt{T} = \sqrt{U + (1 + m^{-1})B},
\]

The total variance \( T \) is corrected in the sense that it is based on the corrected conditional probabilities, and in the sense that not only variability due to measurement error is taken into account, but also variability due to sampling error. According to Rubin, the number of imputations \( m \), can lie somewhere between 2 and 10 to provide reasonable \( T \) values.

Note that only probabilistic assignment rules can be used together with multiple imputation. When deterministic assignment rules, such as modal assignment, are used each individual is assigned to the same latent class in each of the \( m \) repetitions. In this case the between variance \( B \) will be 0.

Table 6 provides the standard errors of the log odds ratios for the relation between \( \theta \) and \( I \) as presented in Table 4 in the one-step procedure, the two-
Table 6: Standard errors of the log odds ratios of Table 4, in the one-step procedure and in the two-step procedure after modal assignment and random assignment, before and after correction. The last two columns contain the log odds ratios after corrected bootstrapped and multi imputed modal and random assignments.

<table>
<thead>
<tr>
<th></th>
<th>1-step</th>
<th>2-step</th>
<th>corr. 2-step</th>
<th>M.I. corr. 2-step</th>
</tr>
</thead>
<tbody>
<tr>
<td>log α</td>
<td>modal</td>
<td>random</td>
<td>random</td>
<td>random</td>
</tr>
<tr>
<td>log α_{12}</td>
<td>0.30</td>
<td>0.16</td>
<td>0.16</td>
<td>0.14</td>
</tr>
<tr>
<td>log α_{13}</td>
<td>0.18</td>
<td>0.13</td>
<td>0.13</td>
<td>0.16</td>
</tr>
<tr>
<td>log α_{23}</td>
<td>0.27</td>
<td>0.14</td>
<td>0.14</td>
<td>0.15</td>
</tr>
</tbody>
</table>

In this paper a comparison is made between one-step and two-step procedures. The two-step procedure, as defined in this paper, has some advantages above the one-step procedure. Instead of one complex model a few smaller models are estimated. In such a way a sparse data solution is created. This makes it possible to handle more variables. The smaller models...
are less prone to local optima, etc. Besides that, it is sometimes impossible to estimate a complex causal model in one step. And sometimes, measurement models are estimated anyway. On the other hand, the one-step procedure has also advantages above the two-step procedure. One of which is very important and first had to be resolved before the two methods can really be compared. In contrast to the one-step procedure, relations between latent variables, or latent variables and external (observed) variables in the two-step procedure are estimated with bias.

In this paper, it is shown why relations between latent variables and external variables are estimated with bias in the two-step procedure, and more important it is shown that it is easy to correct for this bias. Conditional and joint distributions are used to represent the relation between variables. Log odds ratios are used as summarizing measures for the strength of the relation between variables. It is proved that in the two-step procedure in general the log odds ratios of relations between latent and external variables are underestimated. The central idea of the problem is that in the two-step procedure not the relation between the external variables and the latent variables is estimated, but the relation between the external variables and the *predicted* latent variables, which are used as if they are observed. The conditional or joint probabilities of the external variables and the *predicted* latent variables are not the same as the conditional or joint probabilities of the external variables and the latent variables, which are estimated.

However, the conditional (or joint) probability of the predicted latent scores $\tilde{\theta}$ given the external variables $x$, $p(\tilde{\theta}|x)$, can mathematically be written as a function of the true probability $p(\theta|x)$, as follows:

$$E = A.D,$$

where $E$ is the matrix with elements $p(\tilde{\theta}|x)$, $A$ the matrix with elements $p(\theta|x)$, and $D$ is a matrix with known probabilities as elements. In the population case a correction can easily be found with

$$A = E D^{-1},$$

given that $D^{-1}$ exists, which can usually be guaranteed. In practice, when samples are used and elements of $E$ and $A$ are estimated, the formula still can be used because all elements of $D$ are known or estimated after the second step. Of course, the sampling results contain some sampling error, but are generally quite good. The correction formula can be used in all kinds of recursive or non-recursive causal models with one or more external and latent variables. Restrictions on the models are allowed.
In the last part of the paper the one-step procedure and the corrected two-step procedure are compared in a simulation and a real data example. It is shown, among other things that the two-step procedure is less prone to local optima then the one-step procedure, and that the correction of the two-step procedure also works in practice, but then goes together with some errors. Furthermore, some remarks are made about the standard errors. Standard errors of estimated parameters in the two-step procedure are, before but also after correction, under-estimated. This is also due to the fact that in the two-step procedure predicted latent scores are used, while the one-step procedure only considers true latent variables. In the two-step procedure variables are treated as observed while they are in fact latent. This causes underestimation of the variability. A possibility to deal with this can be provided by multiple imputation technique, developed by Rubin (1987).

Appendix A

Theorem 2 is a special case of a theorem, here denoted by Theorem 3, which can be proved by induction.

**Theorem 3** Let $a_i$ and $b_i$ be a series of numbers for which $a_i \geq 0$ and $b_i > 0$, with $i = 1 \ldots n$, then

$$\min_i \frac{a_i}{b_i} \leq \frac{\sum_i a_i}{\sum_i b_i} \leq \max_i \frac{a_i}{b_i}$$

In this Appendix, only the proof (also by induction) of the left part of theorem 2 is presented:

**Proof of the left part of Theorem 2.** Assume that $\min(\frac{u_1}{w_1}, \ldots, \frac{u_n}{w_n}) = \frac{u_m}{w_m}$ and $\max(\frac{u_1}{w_1}, \ldots, \frac{u_n}{w_n}) = \frac{u_M}{w_M}$, and $a_i = u_i v_i$ and $b_i = w_i v_i$, then

i) In the case of $n = 2$:

$$\frac{u_m}{w_m} \leq \frac{u_M}{w_M} \Rightarrow u_m w_M \leq w_m u_M$$
$$\Rightarrow u_m w_M v_M \leq w_m u_M v_M$$
$$\Rightarrow u_m w_M v_M + u_m w_m v_m \leq w_m u_M v_M + u_m w_m v_m$$

$$\Rightarrow u_m (b_m + b_M) < w_m (a_m + a_M)$$
$$\Rightarrow \frac{u_m}{w_m} < \frac{b_1 + b_2}{b_1 + b_2}$$
ii) Suppose for $n = k$: 

$$ \frac{u_{k+1}}{w_{k+1}} \leq \frac{u_m^{(k)}}{w_m^{(k)}} \leq \frac{u_m}{w_m} \leq \sum_{i=1}^{k+1} a_i, $$

then if

$$ a.) \quad \frac{w_{k+1}}{w_{k+1}} \leq \frac{u_{k+1}}{w_{k+1}} \Rightarrow u_{k+1}(\sum_{i}^{k} b_i) \leq w_{k+1}(\sum_{i}^{k} a_i) $$

$$ \Rightarrow u_{k+1}(\sum_{i}^{k} b_i) + u_{k+1}w_{k+1}v_{k+1} \leq w_{k+1}(\sum_{i}^{k} a_i) $$

$$ \Rightarrow u_{k+1}(\sum_{i}^{k+1} b_i) \leq w_{k+1}(\sum_{i}^{k+1} a_i) $$

$$ \Rightarrow \frac{u_{k+1}}{w_{k+1}} \leq \frac{\sum_{i=1}^{k+1} a_i}{\sum_{i=1}^{k+1} b_i}. $$

b.) \quad \frac{u_{k+1}}{w_{k+1}} \geq \frac{u_m^{(k)}}{w_m^{(k)}} \Rightarrow \frac{u_m}{w_m} \leq \frac{u_m^{(k)}}{w_m^{(k)}} \leq \frac{u_m}{w_m}. $$

Furthermore 

$$ u_m^{(k)}(\sum_{i}^{k} b_i) \leq u_m^{(k)}(\sum_{i}^{k} a_i). $$

$$ \Rightarrow u_m^{(k)}(\sum_{i}^{k} b_i) + u_m^{(k)}u_{k+1}w_{k+1}v_{k+1} \leq u_m^{(k)}(\sum_{i}^{k} a_i) + u_m^{(k)}u_{k+1}w_{k+1}v_{k+1} $$

$$ \Rightarrow u_m^{(k)}(\sum_{i}^{k+1} b_i) = u_m^{(k)}(\sum_{i}^{k} b_i) + u_m^{(k)}u_{k+1}w_{k+1}v_{k+1} \leq u_m^{(k)}(\sum_{i}^{k} a_i) + u_m^{(k)}u_{k+1}w_{k+1}v_{k+1} $$

$$ \leq u_m^{(k)}(\sum_{i}^{k+1} a_i) $$

$$ \Rightarrow \frac{u_m^{(k)}}{w_m^{(k)}} \leq \frac{\sum_{i=1}^{k+1} a_i}{\sum_{i=1}^{k+1} b_i}. $$

The proof of the right part goes exactly the same.

\[ \square \]

References


