Methods of Psychological Research Online 2002, Special Issue
Internet: http://www.mpr-online.de
Institute for Science Education
© 2002 IPN Kiel

Questions of a Novice in Latent Markov Modelling*

Frank van de Pol, Statistics Netherlands

Haider Mannan, Canadian Institute for Health Information

1. Introduction

Change is the main issue when panel data are collected. This paper focuses on latent class analysis as a means to represent the regularities in such repeated measurements of the same objects. It does so without adding any novelty to the extensive methodological literature. The emphasis is rather on facilitating the application of these methods. We will elaborate on the questions of a student (Mannan, 2001) who recently finalised a paper on smoking behaviour that relies heavily on latent class analysis.

Before addressing these questions, an overview of the uses of latent class analysis will be given in the next section, followed by a brief introduction to the general model that is used by Mannan for the analysis of his panel data. This model is at the basis of the computer program PANMARK, which the first author developed in collaboration with Rolf Langeheine. In fact, the present paper is written at the occasion of his retirement. And that is a good reason for writing a paper, because the many quotations of his work by various authors show that the views of Rolf Langeheine have shaped the methods for categorical data in discrete time.

---

* The views expressed herein are those of the authors and do not necessarily reflect the policies of Statistics Netherlands or the Canadian Institute for Health Information. Correspondence can be addressed to F.van.de.Pol@home.nl , fpol@cbs.nl and Hmannan@cihi.ca .

2. The use of change models in addition to cross-tables of change

2.1. Overview of functions of latent class analysis

Latent class analysis has many faces. It is a way of looking at categorical data, that is frequency tables. One may use it for data reduction of typically cross-sectional data or for the analysis of repeated measurements, using the assumption of a Markov chain.

Measurement errors and *response uncertainty* are inherent to survey data. They are not only a normal phenomenon in opinion and attitude surveys but also occur when items refer to objective facts such as employment status. Because of these errors, associations are usually underestimated and frequency distributions will be biased, unless errors cancel out. When several indicators of a concept are available, a latent class model enables estimation of misclassification probabilities and also of frequency tables of the 'latent' or 'hidden' variables behind the measured 'indicators' or 'items'. By taking the effect of measurement error into account we get a clearer view on the association between latent variables with generally stronger associations than in the observed frequency tables.

In order to get accurate estimates it is useful, but not always necessary, to have more than one indicator for the same latent concept. When measurements at successive occasions are available, measurement error and 'true' change can be separated, using the assumption of a Markov process.

Population heterogeneity is often shown by cross-classifying a target variable with background-characteristics, like age, sex or region. However, an important part of the variation in the target variable may not be explained by the characteristics that have been measured. The unexplained part may be labelled latent heterogeneity. In the present paper we aim at explaining heterogeneity in change characteristics. A classical example of latent heterogeneity is the difference between people who frequently move to a new house, the 'movers', and people who will not move at all, the 'stayers'. With the

mover-stayer model one can estimate from panel data the proportion movers and the mover's turnover (Blumen, Kogan and McCarthy, 1955). Also more sophisticated mixtures of Markov chains may be estimated, which are used among others in marketing research for modelling brand loyalty (Poulsen, 1990).

For completeness, we also mention *localisation of types*, homogeneous clusters of objects, as a third function of latent class analysis. Using the 'recruitment probabilities', a by-product of latent class analysis, these types may be related to background characteristics. One may for instance measure a dozen sorts of youth criminality. A latent class analysis can reveal types, like 'aggressive behaviour', 'property-motivated behaviour', 'severe criminality' and 'no criminality' and their occurrence. Next, relations between these types and for instance gender or drinking and smoking habits may be analysed.

Reduction of tables is another type of data reduction. In this approach, tables describing a large number of sub-samples are reduced to a small number of latent tables (Clogg and Goodman, 1984; De Leeuw, et al., 1990). Loadings of the sub-samples on these latent tables are generated. For instance a cross-table describing educational mobility may be available for a large number of cohorts. By simultaneous analysis of these cohorts one may estimate for instance two latent tables, one with high mobility and one with low mobility. Older cohorts will load higher on the low mobility table than younger cohorts (Van de Pol, 1997). Instead of a cross-table also a one-way frequency distribution may be analysed, representing for instance time budgets. Apart from this descriptive sort of application, simultaneous analysis of sub-samples may also be used to obtain a maximum likelihood estimate of one table in a panel survey, including all sub-samples that are created by panel attrition.

2.2. Description of data and the models involved

An extensive description of the models involved is given by Langeheine (1994, 2002). The questions that will be addressed in the next section are typical for someone who has read these publications, but lacks application experience. What follows is a brief formal

description of Mixed Markov Latent Class (MMLC) models, based on the user manual of PANMARK (Van de Pol and Langeheine, 1996).

In the simplest case one polytomous variable, x, that is measured at one or more (T) consecutive occasions, $x^1, x^2, \ldots, x^T$, is analysed. A realization in a specific category is denoted i for the first occasion and j, k, ... , m for consecutive time points. This variable which is measured at several occasions can also be viewed as an indicator for a latent variable, also named an indirectly measured or hidden variable. For model identification it is convenient if the development in time of this latent variables is described as a Markov chain. In fact, more than one type of development can be modelled, a mixture of Markov chains. In order to have a well-identified model, also with a latent mixture of Markov chains, another generalisation is useful: more than one indicator can refer to the same latent variable. Finally, a categorical exogenous variable may be introduced by making compartments in the model for several subpopulations.

Analysis with MMLC models focuses on the types of change that exist in these subpopulations. The general MMLC model assumes that each subject belongs to one subpopulation, like gender or birth cohort. Membership of subpopulation h $(h = 1, \ldots H)$ is assumed constant in time, for all indicators. In the sequel, we will refer to manifest, measured, variables as 'indicators', in contrast to the latent 'variables'. The proportion that belongs to subpopulation h is denoted γ_h All other parameters that will be described below are considered conditional on subpopulation h, i.e. all (or some) parameters may be different in each subpopulation.

Each member of subpopulation h belongs to one (latent or manifest) 'chain' of people having the same dynamics. A proportion $\pi_{s|h}$ in subpopulation h belongs to chain s. Hence the proportion in subpopulation h and chain s is $\gamma_h \, \pi_{s|h}$.

A member of subpopulation h and chain s is assumed to belong to one of A classes. The proportion in class a $(a = 1, \ldots A)$ with variable 1, for subpopulation h and chain s,

is denoted. $\delta^1_{a|sh}$ Hence the proportion in subpopulation h, chain s, and class a for variable 1 is $\gamma_h \pi_{s|h}\delta^1_{a|sh}$.

The probability to answer i for indicator 1, given h, s and a, the response probability $\rho^{1\,1}_{i|ash}$, is assumed the same for all subjects in subpopulation h, chain s and class a. Hence the proportion in subpopulation h, chain s, class a with variable 1, and category i for indicator 1 is $\gamma_h \pi_{s|h}\ \delta^1_{a|sh}\ \rho^{1\,1}_{i|ash}$. If the variables of the model are not latent but manifest, the latent classes, a for variable 1, b for variable 2, etc., coincide with the manifest categories, i for indicator 1, j for indicator 2, etc.. Moreover, the response probabilities $\rho^{1\,1}_{i|ash}$ are superfluous in a manifest model: $\rho^{1\,1}_{i|ash} = 1$ for a $a = i$ and else 0.

If the model does not only involve a latent variable at time 1, but also one at time 2, then, for subpopulation h, each member of chain s is assumed to behave according to the same transition probabilities, $\tau^{2\,1}_{b|ash}$, from class α for time 1 to class b for time 2. As for the time 1 indicator, also for the time 2 indicator the probability to answer j, given being in class b, chain s and subpopulation h, $\rho^{2\,2}_{j|bsh}$, is assumed to be the same for all subjects in subpopulation h, chain s and class b. Hence P_{hsaibj}, the proportion in subpopulation h, chain s, class a for latent variable 1, category i for indicator 1, class b for latent variable 2 and category j for indicator 2, is $\gamma_h\ \pi_{s|h}\ \delta^1_{a|sh}\ \rho^{1\,1}_{i|ash}\ \tau^{2\,1}_{b|ash}\ \rho^{2\,2}_{j|bsh}$.

An expression for the observable population table, P_{hij}, of two manifest indicators (not the latent variables), is obtained by summing over the latent indices s, a and b. The model equation for the latent mixed Markov chain for three time points is written as

$$P_{hijk} = \gamma_h \sum_{s=1}^{S}\sum_{a=1}^{A}\sum_{b=1}^{B}\sum_{c=1}^{C} \pi_{s|h}\ \delta^1_{a|sh}\ \rho^{1\,1}_{i|ash}\ \tau^{2\,1}_{b|ash}\ \rho^{2\,2}_{j|bsh}\ \tau^{3\,2}_{c|bsh}\ \rho^{3\,3}_{k|bsh}$$

Extension to multiple indicators is straightforward (Langeheine and Van de Pol, 1993,1994). For two indicators at each time point, with subscripts i and i' for time point 1, subscripts j and j' for time point 2, subscripts k and k' for time point 3, we would get

$$P_{hii'jj'kk'} = \gamma_h \sum_{s=1}^{S} \sum_{a=1}^{A} \sum_{b=1}^{B} \sum_{c=1}^{C} \pi_{s|h} \, \delta^{1}_{a|sh} \, \rho^{1\,1}_{i|ash} \, \rho^{2\,1}_{i'|ash} \, \tau^{2\,1}_{b|ash} \, \rho^{3\,2}_{j|bsh} \, \rho^{4\,2}_{j'|bsh} \, \tau^{3\,2}_{c|bsh} \, \rho^{5\,3}_{k|csh} \, \rho^{6\,3}_{k'|csh}$$

In a (panel) sample an estimate, p_{hijk} is found or, in the case of two indicators for each occasion, $p_{hii'jj'kk'}$. From these estimates one can compute maximum likelihood parameter estimates for the γ, π, δ, ρ and τ, for instance with the computer program PANMARK, provided that identifying restrictions are applied. It uses a version of the EM algorithm (Van de Pol and De Leeuw, 1986; Van de Pol and Langeheine, 1990).

The present notation in terms of conditional probabilities can be replaced by a notation in terms of log-linear parameters (Vermunt, 2001). An advantage of the latter notation is that it allows generalization to more complex models. On the other hand, probabilities that are on the boundary of the parameter space, 0 or 1, are not easily estimated in this approach since they correspond to log-linear parameters that go to infinity, and computers set a limit to any number.

3. Questions and answers

3.1. Getting started

A scholar in Canada, named Haider Mannan, is preparing a paper to become a Master of Science in Epidemiology and Biostatistics and decides to use latent Markov models. He obtains PANMARK and begins to pose questions. At first some trivial questions arose, like

"How can I manipulate a Windows shortcut to a DOS program?".

When a program or its working directory are moved to a non-default place, a shortcut to the program does not function anymore. Experienced Windows users know that they need to check the shortcut's properties, especially the program-tab, where the first line is for starting the exe-file and where the second line should contain the directory where that program will start reading and writing files, the working directory. Next comes a phase, when the computer program is unfamiliar.

I tried to fit a one-chain manifest Markov model (both transition and response probabilities time homogeneous) by creating a restriction file and starting values file, but the PANMARK output was wrong ?? In Model Definition I selected 'Latent Markov chain' since there is no manifest Markov chain option.

Specifying restrictions for a large number of parameters can easily be mixed up. The present program identifies each parameter from the order in which it appears. The user can copy this order from the file with parameter estimates, but apparently this is an error prone procedure. However, a single manifest Markov chain can also be estimated in a much simpler way, that is by selecting a standard analysis: a mixed Markov model with only 1 chain. This does not require any extra input with respect to restrictions or starting values.

Soon the use of starting values becomes inevitable for the present data, because some transitions are impossible at the latent level. In an educational setting such models involve irreversible steps in learning (Collins et al., 1992; Langeheine et al., 1994).

"I have smoking data for grade 6, grade 8 and grade 11 students. ..." "... My supervisor wants me to fit the models by restricting three parameters to 0, which are transition probabilities from state current smoker (1) to never smoker(4), from ex-smoker (2) to never smoker, and from experimental smoker(3) to never smoker. ..." "...I did not get the correct result for ... latent Markov model when I fixed 3 transition probabilities (from state 1 to 4, 2 to 4, and 3 to 4) ... to 0."

Transition from any of the smoking conditions to "never smoked" (4) is impossible if people make no errors in responding. The latent Markov model can be used to separate response errors from "true change". Doing so, the latent matrix of transition probabilities should have zeroes for these impossible transitions, i.e. $\tau^{86}_{4|a} = 0, \tau^{118}_{4|b} = 0\,(a,b \neq 4)$, with superscripts indicating the occasions: grades 6, 8 and 11. Mannan had taken a file with estimates and altered these near-zero parameters into exactly zero. Starting the estimation of a parameter with 0 has the same effect as fixing it to 0, because the present program uses the EM algorithm. Iterations in this algorithm involve multiplicative adjustment of estimates, thus leaving zeroes at 0 until convergence.

Mannan got confused, because the program warned that "Row ending" <some number> "does not add up to 1". If someone is in smoking state 2, for instance, he or she must show up in any of the possible states next time (supposing there was no drop out). This property, summation to one, holds for any set of conditional probabilities, be it $\rho^{t}_{i|a}$, response probabilities given latent state a, π_s proportions belonging to any chain $s = 1, \ldots, S$ or $\tau^{86}_{j|i}$ manifest transitions originating from one original status, i. If one probability in a set is fixed to zero, the starting values for the remaining probabilities of the same "row" must be increased to the extent that they add up to 1.

3.2. Time homogeneity

Stationarity or *time homogeneity* is a recurring issue in latent Markov models. It is the assumption that the probability to move from state a to state b is the same for all time intervals. It is attractive to make this assumption, since the accuracy of the estimates is much better if the assumption is correct, i.e. standard errors are smaller. For Mannan's data this assumption is questionable.

> *"(The) Latent Markov model may require equally spaced measurements for fitting a totally constrained model .... Since my smoking measurements are unequally spaced (in time) do you think that it will be inappropriate to fit the above mentioned model?"*

Possibly more turn-over between latent states takes place in a longer time interval. Therefore the off-diagonal transition probabilities, $\tau^{86}_{b|a}(b \neq a)$, for the two-year period should be allowed to be smaller than the corresponding ones in the three year period that followed. Of course estimates will deviate somewhat from this expectation due to sampling error and as a consequence some off-diagonal probabilities will suggest more change in the shorter period, contrary to our expectation. Suppose the only exception is $\hat{\tau}^{86}_{3|4} > \hat{\tau}^{118}_{3|4}$, then I would suggest to re-estimate the model with assumption $\tau^{86}_{3|4} = \tau^{118}_{3|4}$.

3.3. Restrictive mixed Markov models

In his search for a good fitting model, Mannan tested among other things whether part of the respondents responded randomly.

> *It is not clear to me how black-and-white and mover-random response models can be fitted by* PANMARK*. The manual does not give examples of these models.*

In a "*random response*" chain the all transition probabilities should be fixed to 0.25 in the present case of four response categories. Less restrictive and well known from log-linear models is a model that assumes independence between occasions. In the present context *independence* is only assumed for one "chain". It means that for instance $\tau^{8\,6}_{j|i} = \tau^{8\,6}_{j}$, irrespective of the originating category i. Denoting transition probabilities that are equal with the same number, these equality restrictions are written

1 2 3 4
1 2 3 4
1 2 3 4
1 2 3 4,

with rows indicating grade 6 category and columns the grade 8 category.

3.4. Second order Markov models

Mannan also fitted a second-order Markov chain. This models allows for the initial, grade 6, situation making a difference for the change between the two occasions that followed, from grade 8 to rade 11.

> *Could you please explain the rules for arranging frequencies in the dataset for 2nd and higher order Markov models ? I don't understand when and how zero frequencies appear.*

In among others Langeheine and Van de Pol (2000) frequency data are arranged in such a way that a program for first-order Markov chains can fit second order Markov chains. The trick is that each cross-table of consecutive measurements, for instance occasion 1 times occasion 2, is presented to the program as one variable. The observed frequencies are interspersed with zero frequencies at positions that are impossible, like 1

and 1 at occasions 1 and 2 respectively in combination with 2 and 1 at occasions 2 and 3 respectively. Occasion 2 cannot have score 1 and 2 at the same time.

With data on three occasions only, the second order Markov chain is not a restrictive model. It is merely a reparametrization of the data, a saturated model. For testing the fit of mixed Markov latent class models one needs two or more indicators at each occasion (Langeheine and Van de Pol, 1990). Measurements on more than three occasions are also helpful. Mannan laid his hands on measurements on four occasions: grades 6, 8, 10 and 12.

With data on four occasions Langeheine's *partially latent mixed Markov* model could be estimated (Langeheine and Van de Pol, 1992). This is a mover-stayer model with response error in the mover type only. The assumption that stayers don't make response errors is both plausible and practical. Plausible because it is more easy to produce the correct answer if one's position is stable. Practical because data on one indicator measured at four occasions are not informative enough to estimate a more complex model.

> *Hope you are well. Presently I am analyzing the second and last data set. There are 256 = 4×4×4×4 cells (i.e. there are 4 indicators (smoking) each having 4 categories). Many of the cells are 0. I successfully fitted manifest Markov and latent Markov model with stationary and non-stationary transition probabilities and response probabilities.... But, for latent mover-stayer model with time-dependent transition probabilities, all parameters are not identified. This model has considerably more degrees of freedom than the # of parameters. I tightened the stopping criterion to 10^{10}. But, still all parameters are not identified. I don't understand why this is happening. I fixed the response probabilities. for stayers to 0 and 1, but still all parameters are not identified. ...*

Fitting a latent mover stayer model generally is asking too much from the data. Even if you have four or more panel waves, standard errors are very large and for some datasets the algorithm won't even converge. Therefore I have assumed that the stayers know what they are talking about, i.e. the reliability of their answers is perfect: the partially latent mover stayer model as Rolf Langeheine calls it. Because you don't want to assume stationary transition probabilities, even that restricted PLMS model has some pa-

rameters that are hard to estimate. While running the analysis, a first indication of this is the slow convergence of the algorithm. For each iteration, only a small improvement in fit is obtained. This usually means that some parameters are highly correlated and hence their standard deviations are large. Sometimes this can only be mended by adding some restrictions to the model.

For your dataset not all parameters of the PLMS model can be identified. This can be seen from the information matrix, which is not positive definite. Not all Eigenvalues are positive. When you take a look at your data, you will see that only 9 respondents remain in category 4, "never smoker", and few respondents leave this condition. So there is almost no information to separate reliability and stability for this group. We have to make more assumptions. Would it be reasonable to assume that class 3 (experimenting with smoking) is empty with stayers? After all, you cannot continue experimenting for ever. It does not make sense to be experimenting during 6 years, especially because stayers are assumed to respond with perfect reliability. Because only one eigenvalue is 0, I have tried to estimate this model.

As the output shows, you will have to make more assumptions in addition. I suggest you make some stationarity assumption. You don't have to assume the complete transition parameter matrices (t1xt2, t2xt3 and t3xt4) to be equal. It may be enough to set only some of the content of these matrices equal. I cannot say which parameters can safely be assumed to be stationary, without messing up your main hypothesis. (I suppose you want to assess the effect of some experiment between time points.)

Another thing to keep in mind is that the first wave of a panel often is measured a little less reliably than the occasions that follow. If you want to assess this phenomenon by allowing the response probabilities of the first occasion to be different, you will have to assume stationary transition probabilities.

Finally Mannan tried to fit a latent second-order Markov chain.

"...I have reanalyzed 2nd order Markov model by creating a totally different starting values file which I gave you. The transition matrix which I created in this file has 16 rows and 16 columns. But still I am not getting the correct result ..."

Mannan forgot to include a line with delta's in the starting values file. So there is no initial distribution defined. Moreover, there is another problem. Despite the fact that PANMARK says it can handle more than 200 categories, in fact the limit is 15 at the moment. The problem is that internally a 16x16 transition matrix has 256 entries and that is just one too many for the routine that sets equality restrictions.

3.5. The bootstrap for assessing model fit

The likelihood ratio is a model fit criterion that can be evaluated with the chi-square distribution as long as sample size is large enough in every cell of the contingency table. Another requirement for using this theoretical distribution is that there is no doubt about the degrees of freedom. Sometimes parameters get fixed during estimation to one of the bounds of the parameter space, 0 or 1. The effect of such boundary values on the degrees of freedom is uncertain. Finally, equality restrictions do strictly spoken prohibit the use of the simple theoretical chi-square distribution, since a more complicated distribution applies with equalities.

If one of these conditions is not met, the distribution of the fit statistic can be generated with repeated parametric bootstrap samples (Langeheine et al., 1996). Now that fast computers offer the computational possibilities, researchers will test the fit of a model using the bootstrap, even if sample size is large.

Hope you received my last email. I am having problems with bootstrap analysis. I ran a latent Markov model with unrestricted parameters last night at around 9 and until 2.30 am today the run was still going on (i.e., for 17.5 hours) with only 200 bootstrap samples being drawn. The data set has only 64 frequencies with 16 being zero. So I really lost my patience and stopped the run. The stopping criterion was set at 1e-8 (the default). It is likely that the algorithm for the program is not efficient or the algorithm is spinning because of the tight criterion for convergence. In case the algorithm spins, then a larger criterion should be used. What stopping criterion do you recommend ? For mixed Markov models including mo-

ver-stayer, black-and-white etc. what stopping criterion do you suggest ? Note that for my run I selected the option "the EM algorithm will stop if bootstrap LR is less than the original LR".

If the algorithm seems to be spinning, there is an identification problem. Either the model you specified is not identified for these data, or the likelihood surface is almost flat, thus making it almost impossible to find the optimum. In the first case, one or more eigenvalues will be zero or negative. In the second case, one or more eigenvalues will be very small in comparison to the biggest one. In the first case you cannot compute standard errors of the parameters, in the second case some standard errors will be very large. The conclusion is that you should always check the identification of your model before asking for a bootstrap analysis. The model should be identified and standard errors should be reasonably low. After all, what can we conclude from an analysis that gives model parameters that are very inaccurately determined? Finally, even if the model is identified for the original sample, you may have a special case that generates some bootstrap samples for which themodel is not well identified.

Moreover, the option to have a large number of randomly chosen starting values is too time consuming to be applied in combination with bootstrapping.[1] Next best is to have estimation start with the estimates from the original sample.

Since there are fewer cells than cases I selected 'sampling from cells'.

There are two ways to draw a non-naïve, parametric bootstrap sample. The most common approach is to draw a sample from the multinomial distribution with the population proportions that were estimated after fitting the model. In PANMARK this is

[1] Latent class models may have more than one optimum in the likelihood surface. Of course, the most likely one is to be preferred, but the problem is that an optimisation routine may end up in a local optimum. Therefore it is recommended to start the algorithm with several iterations from many points of the likelihood surface and to continue with the most promising locations. The more parameters are involved, the more starting points are needed. The relationship is not linear, but seems rather exponential. Doubling the number of independent parameters one should take the square number of starting points. In practice, this is not feasible, though.

implemented by drawing successive binomial samples and referred to as "sampling from cells". For models that can be written in terms of conditional probabilities there is an alternative approach, that is based on drawing each case according to the model probabilities. This is referred to as sampling from cases (Langeheine et al., 1996).

Sampling from cases is to be preferred always! Sampling from cells has a (small) bias in the last cell, I found recently, at least in my implementation. I will eliminate the sampling from cells option from new program versions.

3.6. Exogeneous variables, covariates

At the beginning of his analysis Mannan had the intention to explain latent parameters like the proportion who stopped smoking.

> *... For analysing a covariate do we only need the covariate being measured at the first time point, say the covariate being measured at grade 6 for transitions of all students from grade 6 to 12, is sufficient for analysing such a covariate. If that is so, then we are assuming that the covariate is time independent. ...Can we handle a time dependent covariate and how ? ...Can we handle continuous covariates in latent Markov models in any way?*

There are two schools of thought with respect to incorporating explanatory, exogenous variables into a measurement model. One school emphasizes the importance of simultaneous estimation of all parameters involved (Van der Heijden et al., 1996; Muthen, 2002). On the other hand it has been argued that explanatory variables should not have any influence on the measurement model. An explanatory variable that associates strongly with only one of the indicators in the measurement model will easily give an unrealistic boost to the reliability estimate of that indicator and the reliability of the other indicators will seems smaller than they really are.

In the latter approach you can use the recruitment probabilities (which are almost the same as grades of membership, GOM's) to get information on class membership for

every response pattern that you have in your sample.[2] This information is a probability distribution of latent variables for each combination of the indicators, $\xi^{681012681012}_{a\,b\,c\,d\,|\,i\,j\,k\,l}$. In the presence of a latent table with (theoretically) 4^4=256 cells, you will probably aggregate the $\xi^{681012681012}_{a\,b\,c\,d\,|\,i\,j\,k\,l}$ by summation over most of the latent variables, focussing on for instance the latent proportions at occasion 2.[3] Let us have a look at two score patterns and the accompanying recruitment probabilities for class $b = 1, 2, 3$ and 4 respectively:

Score pattern 1, 1, 2, 2: recruitment probabilities .2 .1 .4 .3

Score pattern 2, 2, 1, 1: recruitment probabilities .4 .3 .2 .1

Next, you create a new variable in your data file, called "latent class". Make four copies of your data file, with latent class = 1, 2, 3 and 4 respectively. Stack these four files below each other in one new file. This new file has four times as many cases as the original one. In order to match this new data file to these recruitment probabilities they should be rewritten as follows.

Score pattern 1, 1, 2, 2, class 1: recruitment probability: .2

Score pattern 1, 1, 2, 2, class 2: recruitment probability: .1

Score pattern 1, 1, 2, 2, class 3: recruitment probability: .4

Score pattern 1, 1, 2, 2, class 4: recruitment probability: .3

Score pattern 2, 2, 1, 1, class 1: recruitment probability: .4

Score pattern 2, 2, 1, 1, class 2: recruitment probability: .3

Score pattern 2, 2, 1, 1, class 3: recruitment probability: .2

Score pattern 2, 2, 1, 1, class 4: recruitment probability: .1

---

[2] Don't use multiple groups for this with older versions of PANMARK. The GOM's are only correct with 1 group.

[3] A bit more complicated is to focus on the transition from smoker to ex-smoker. This would involve summation of the $\xi^{681012\;681012}_{a\,b\,c\,d\,|\,i\,j\,k\,l}$ over latent indices c and d. In order to reduce the size of the latent cross-table, also the number of latent states at occasions 1 and 2 should be reduced.

And so on for all score patterns in the analysis. Each record represents (partial) membership of one category of the latent variable at occasion 2. Then the fourfold data file can be matched with these recruitment probabilities, using the score pattern and the latent class as matching keys. Finally, the recruitment probabilities should be used as weights in the analyses that follow. Note that every case still has weight one, being the sum of all four recruitment probabilities of the relevant score pattern. With the resulting data file you can examine membership of a specific cell of the latent variable in a logit or probit analysis with as many covariates as you like.

The advocates of simultaneous estimation, full information maximum likelihood (FIML), can argue that there is no evidence the measurement model holds for the whole population. If for instance reliability is lower for male smokers than for female smokers and this difference is ignored in the measurement model, the influence of gender on smoking behaviour will be estimated with a bias. This type of heterogeneity can be tested with simultaneous estimation of the measurement model in subgroups of the sample. However, sample size sets limits to this type of corroboration of results.

4. Conclusion

The readers who have accompanied us to this point might want see an example. For this, we can refer to the literature, in the present volume for instance. Or better even, we should refer to the reader's own data that can be analysed with latent class models.

References

[1] Blumen, I., Kogan, M. & McCarthy, P.J. (1966). Probability models for mobility. In P.F. Lazarsfeld & N.W. Henry (eds.), *Readings in mathematical social science*, pp.318-334. reprint from: The industrial mobility of labor as a probability process, Cornell Univ. Press, Ithaca, N.Y., idem., 1955. Cambridge, Massachusetts: MIT Press.

[2] Clogg, C.C. & Goodman, L.A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association 79*, pp. 762-771.

[3] Collins, L.M. & Wugalter, S.E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research 27*, pp. 131-157.

[4] De Leeuw, J., van der Heijden, P.G.M. & Verboon, P. (1990). A latent time-budget model. *Statistica Neerlandica 44*, pp. 1-22.

[5] Langeheine, R. (1994). Latent Variable Markov Models. In A. von Eye & C.C. Clogg (eds.), *Latent Variables Analysis. Applications for Developmental Research*, pp. 373-395. Thousand Oaks, California: Sage.

[6] Langeheine, R. (2002). Latent Markov Chains. In A.L. McCutcheon & J.A. Hagenaars (eds.), *Advances in Latent Class Analysis*. Cambridge: University Press.

[7] Langeheine, R. & van de Pol, F. (1990). A unifying framework for Markov modeling in discrete space and discrete time. *Sociological Methods & Research 18*, pp. 416-441.

[8] Langeheine, R. & van de Pol, F. (1993). Multiple Indicator Markov models. In R. Steyer, K.F. Wender & K.F. Widaman (eds.), *Psychometric Methodology. Proceedings of the 7th European Meeting of the Psychometric Society in Trier 1993*, pp. 248-252. Stuttgart, New York: Fischer.

[9] Langeheine, R. & van de Pol, F. (1994). Discrete time mixed Markov latent class models. In A. Dale & R. Davies (eds.), *Analyzing social & political change: a Casebook of Methods*, pp. 167-197. London: Sage.

[10] Langeheine, R., Stern, E. & van de Pol, F. (1994). 'State mastery learning: dynamic models for longitudinal data'. *Applied Psychological Measurement 18*, pp. 277-291.

[11] Langeheine, R., Pannekoek, J. & van de Pol, F. (1996). Bootstrapping Goodness-of-Fit Measures in Categorical Data Analysis. *Sociological Methods & Research 24*, pp. 492-516.

[12] Langeheine, R. & van de Pol, F. (2000). Fitting Higher Order Markov Chains. *Methods of Psychological Research 5,* www.mpr-online.de , Pabst Science Publishers.

[13] Mannan, H.R. (2001). *Latent Markov Modelling of Smoking Transitions.* London, Canada: University of Western Ontario.

[14] Muthen, B. (2002). *Beyond SEM: General Latent Variable Modeling.* http://www.statmodel.com/Behaviormetrikapaper1.pdf.

[15] Poulsen, C.S. (1990). Mixed Markov and latent Markov modelling applied to brand choice behaviour. *International Journal of Research in Marketing 7*, pp. 5-19.

[16] Van der Heijden, P., Dessens, J. & Böckenholt, U. (1996). Estimating the Concomitant Variable Latent Class Model with the EM Algorithm. *Journal of Educational and Behavioral Statistics 21*, pp. 215-229.

[17] Van der Heijden, P., 't Hart, H. & Dessens, J. (1997). A Parametric Bootstrap Procedure to Perform Statistical Tests in a LCA of Anti-Social Behaviour. In J. Rost & R. Langeheine (eds.), *Applications of Latent Trait and Latent Class Models in the Social Sciences*, pp. 196-208. Münster: Waxmann.

[18] Van de Pol, F.J.R. & de Leeuw, J. (1986). A latent Markov model to correct for measurement error. *Sociological Methods & Research 15*, pp. 118-141.

[19] Van de Pol, F.J.R. & Langeheine, R. (1990). Mixed Markov latent class models. In C.C. Clogg (ed.), *Sociological Methodology*, pp. 213-247. Oxford: Blackwell.

[20] Van de Pol, F.J.R., Langeheine, R. & de Jong, W.A.M. (1996). *PANMARK 3 user's manual; PANel analysis using MARKov chains, a Latent Class Analysis program.* F.van.de.Pol@home.nl.

[21] Van de Pol, F.J.R. (1997). Educational mobility, cohort and gender, a latent class re-analysis of the Ganzeboom and De Graaf data. In J. Rost & R. Langeheine (eds.), *Applications of latent trait and latent class models in the social sciences*, pp. 412-419. Münster/New York: Waxmann.

[22] Vermunt J.K., Rodrigo, M.F. & Ato-Garcia, M. (2001). *Modeling Joint and Marginal Distributions in the Analysis of Categorical Panel Data.*