Statistical Methods in Medical Research

Latent class and finite mixture models for multilevel data sets

Jeroen K Vermunt

The online version of this article can be found at:

Additional services and information for *Statistical Methods in Medical Research* can be found at:

Email Alerts: http://smm.sagepub.com/cgi/alerts

Subscriptions: http://smm.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations: http://smm.sagepub.com/content/17/1/33.refs.html

>> Version of Record - Mar 14, 2008

OnlineFirst Version of Record - Sep 13, 2007

What is This?

Latent class and finite mixture models for multilevel data sets

Jeroen K Vermunt Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands

An extension of latent class (LC) and finite mixture models is described for the analysis of hierarchical data sets. As is typical in multilevel analysis, the dependence between lower-level units within higher-level units is dealt with by assuming that certain model parameters differ randomly across higher-level observations. One of the special cases is an LC model in which group-level differences in the logit of belonging to a particular LC are captured with continuous random effects. Other variants are obtained by including random effects in the model for the response variables rather than for the LCs. The variant that receives most attention in this article is an LC model with discrete random effects: higher-level units are clustered based on the likelihood of their members belonging to the various LCs. This yields a model with mixture distributions at two levels, namely at the group and the subject level. This model is illustrated with three rather different empirical examples. The appendix describes an adapted version of the expectation–maximization algorithm that can be used for maximum likelihood estimation, as well as providing setups for estimating the multilevel LC model with generally available software.

1 Introduction

Latent class (LC) analysis is becoming an increasingly popular tool for building diagnostic classifications and determining the sensitivity and specificity of diagnostic measures,[1] for determining the agreement between raters,[2] and for regression modelling with multiple discrete outcome variables.[3,4] Other more exploratory application types include scaling,[5–7] clustering,[8–10] and nonparametric random-effects modelling.[11–13] LC analysis belongs to the family of latent variable methods: a discrete unobserved variable with K categories is assumed to exist, which affects the, usually multiple, responses of subjects. Each subject is assumed to belong to one of these K LCs. Statisticians often refer to the methods discussed in this article using the more technical term finite mixture models and reserve the term LC analysis for the situation in which all response variables are categorical.[8,9] Here, the two terms will be used interchangeably.

In medical research, data sets that show a hierarchical or multilevel structure are often encountered.[12,14,15] Examples are data on patients who are nested within centres, respondents nested within regions, children nested within families and repeated measures nested within subjects. Also when applying LC analysis, the data at hand may show this type of nesting, which has to be taken into account when modelling the data.

Address for correspondence: Jeroen K Vermunt, Department of Methodology and Statistics, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: j.k.vermunt@uvt.nl

Table 1 Four-fold classification of multilevel latent variable models

Subject-level (level 2) latent variable(s)	Group-level (level 3) latent variable(s)	
	Discrete	Continuous
Discrete	I. Multilevel mixture LC	II. Multilevel random-effects LC
Continuous	III. Multilevel mixture IRT/FA	IV. Multilevel random-effects IRT/FA

This article presents an extension of LC analysis that can be used to deal with such multilevel data structures. As demonstrated by Vermunt,[4,16] this can be achieved by introducing either discrete or continuous random effects in the model of interest; that is, by introducing one or more discrete or continuous latent variables at the higher level of the hierarchy.

It should be noted that this multilevel LC model is actually a model for three-level data since the LC model itself is already a two-level model, that is, a model for multiple responses per subject. The presented extension of the LC model can therefore be seen as a special case of a more general family of latent variable or random-effects models for three-level data sets, which are generalized linear models containing discrete and/or continuous random effects (or latent variables) at levels 2 and 3.[12] These methods can be classified using the four-fold classification presented in Table 1.

Model types I and II are LC or finite mixture models with either discrete or continuous random effects at level 3, which are the models that will be discussed in more detail in this article. Types III and IV are multilevel variants of continuous latent variable models such as factor analysis (FA) and item response theory (IRT) models. Various special cases of type IV have been presented in the literature.[17–21] The very interesting model type III – a continuous latent variable model with a discrete mixture at level 3 – is explored in more detail in an accompanying article.[22] This full framework is implemented in the user-friendly Latent GOLD computer program for latent variable modelling[23] and thus readily available for any interested researcher. In addition to these four special cases, hybrid variants can be defined combining discrete and continuous latent variables at both levels of the hierarchy.[24]

The next section introduces LC analysis for nonhierarchical data. Subsequently, the multilevel extension of the LC model, which involves including random effects in certain parts of the model of interest, is discussed. The approach is illustrated with various empirical examples. In the appendix technical details on parameter estimation by maximum likelihood (ML), as well as syntax files for running the empirical examples with generally available software, are provided.

2 LC and finite mixture models

Let y_{it} denote the response of subject i on the tth response variable, and the corresponding vector of responses by $\mathbf{y}_i$. These multiple responses may be different measures or symptoms of the same disease, ratings of different raters, repeated measures on the same individuals in a longitudinal study or other types of multiple observations. The number

of observations on a subject is denoted by T, $1 \leq t \leq T$. In applications in which the number of observations may vary across subjects, this number is denoted by T_i instead of T.

The basic idea of an LC or finite mixture model is that subjects belong to one of K LCs and that the multiple responses of a subject are generated by class-specific densities or probabilities.[9] The underlying discrete latent variable will be denoted by x_i and a particular value (a particular LC) by k, $1 \leq x_i = k \leq K$. The basic formula of an LC or finite mixture model is the following

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} P(x_i = k) f(\mathbf{y}_i | x_i = k)$$

$$= \sum_{k=1}^{K} P(x_i = k) f(\mathbf{y}_i; \boldsymbol{\vartheta}_k) \tag{1}$$

The density corresponding to the responses of subject i, $f(\mathbf{y}_i)$, is assumed to be a mixture of class-specific densities $f(\mathbf{y}_i; \boldsymbol{\vartheta}_k)$, where the mixture probabilities $P(x_i = k)$ serve as weights and $\boldsymbol{\vartheta}_k$ are the parameters defining the class-specific densities. The choice of the specific form for $f(\mathbf{y}_i; \boldsymbol{\vartheta}_k)$ depends of the scale type of the response variables. With continuous responses, one may, for example, define the class-specific densities to be multivariate normal.[9,10]

An additional assumption that is often made is that the T responses are conditionally independent of one another given class membership of subject i, which is sometimes referred to as the local independence assumption.[5,6] This implies that

$$f(\mathbf{y}_i; \boldsymbol{\vartheta}_k) = \prod_{t=1}^{T} f(y_{it}; \boldsymbol{\vartheta}_{tk}) \tag{2}$$

where $\boldsymbol{\vartheta}_{tk}$ are the parameters defining the distribution of response variable t in LC k. As was indicated above, in certain applications, T may be replaced by T_i to allow for different numbers of responses per subject. Combining these two basic Equations (1) and (2) yields

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} P(x_i = k) \prod_{t=1}^{T} f(y_{it} | x_i = k)$$

$$= \sum_{k=1}^{K} P(x_i = k) \prod_{t=1}^{T} f(y_{it}; \boldsymbol{\vartheta}_{tk})$$

Note that the local independence assumption, which may be (partially) relaxed,[25] is typical for any type of latent variable and two-level regression model, and thus not

specific for LC analysis. This assumption not only simplifies the form of the class-specific densities to univariate densities, such as Poisson for counts, (single trial) multinomial for discrete responses and univariate normal for continuous responses, but may also be required for identification. More specifically, with dichotomous and nominal responses, multiple locally independent response variables are needed in order to be able to identify the parameters of an LC model. In contrast, with counts and continuous variables, a single response per subject suffices, though, of course, more interesting models may be built with multiple responses, in which case the local independence assumption is not required for identification.

In these examples, solely local independence and single response variable models will be dealt with, which means that the class-specific densities are univariate distributions, where the exact form depends on the scale type of the response variable concerned. Note that it is also possible to deal with mixed scale types, in which case different types of densities are combined in the same model.[8] In the examples, three types of densities, Poisson, univariate normal and (single trial) multinomial, are used yielding

$$f(y_{it}; \theta_{tk}) = \frac{\theta_{tk}^{y_{it}}}{y_{it}!} \exp(-\theta_{tk})$$

$$f(y_{it}; \mu_{tk}, \sigma_{tk}^2) = \frac{1}{\sqrt{2\pi\sigma_{tk}^2}} \exp\left\{-\frac{1}{2}\frac{(y_{it} - \mu_{tk})^2}{\sigma_{tk}^2}\right\}$$

$$f(y_{it}; \boldsymbol{\pi}_{tk}) = \prod_{m=1}^{M} (\pi_{tkm})^{\delta_{itm}}$$

respectively, where $\delta_{itm} = 1$ if $y_{it} = m$ and 0 otherwise and M is the number of categories in the multinomial case. The model parameters for response variable t and LC k are the mean θ_{tk} in the Poisson model, the mean μ_{tk} and variance σ_{tk}^2 in univariate normal model, and the probabilities π_{tkm} associated with the M categories – which can be collected in the vector $\boldsymbol{\pi}_{tk}$ – in the multinomial model. In the normal model, it is rather common to restrict σ_{tk}^2 to be equal across LCs ($\sigma_{tk}^2 = \sigma_t^2$), which amounts to assuming homoscedastic errors. Note that in the dichotomous case ($M = 2$) the multinomial model simplifies to a Bernoulli model.

An important extension of the simple LC and finite mixture models described above is obtained by including predictors in the model for the LCs[3,26] and/or for the response variables.[3,11–13] This involves specifying a multinomial logistic regression models for x_i or a regression model from the generalized linear modelling (GLM) family for y_{it}, where for ordinal categorical response variables, an ordinal regression model may be used. Other extensions and variants include models with multiple discrete latent variables,[7] models with local dependencies[25] and models that contain continuous latent variables in addition to the discrete latent variable.[27,28]

3 Multilevel extension

For the description of the multilevel LC model, the notation introduced in the previous section has to be expanded somewhat. The nested data structure implies that subjects belong to one of J groups, where a particular group is referred to by the index j, $1 \leq j \leq J$, and the number of subjects in group j by n_j. The group variable will be denoted by G. A particular response variable, the vector of T response variables and the class membership of subject i in group j are now denoted by y_{jit}, $\mathbf{y}_{ji}$ and x_{ji}, respectively.

Multilevel analysis implies that the parameters of the model of interest are allowed to differ across groups. The same applies for the multilevel variant of the LC model. In the fixed effects approach, a multilevel LC model takes on the form of a multiple-group LC model

$$f(\mathbf{y}_{ji}|G=j) = \sum_{k=1}^{K} P(x_{ji}=k|G=j) \prod_{t=1}^{T} f(y_{jit}|G=j)$$

$$= \sum_{k=1}^{K} P(x_{ji}=k|G=j) \prod_{t=1}^{T} f(y_{jit}; \boldsymbol{\vartheta}_{jtk})$$

that is, a model in which each group has its own set of class membership probabilities and $\boldsymbol{\vartheta}$ parameters. It will be clear that such a model cannot be estimated in a typical multilevel application, in which there are more than a few groups – say 50 or more – and in which the number of subjects per group is rather small – say 30 or less.

Typical for multilevel analysis is that the group differences are dealt with by means of random instead of fixed effects.[14,15,29,30] A multilevel LC model is thus obtained by introducing one or more group-level continuous random effects – continuous latent variables – in the model to account for the fact that some of the LC model parameters differ across groups. An alternative is to introduce a group-level discrete latent variable in the model, where parameters are allowed to differ across LCs of groups.[16] This rather flexible approach is similar to nonparametric random-effects modelling.[11–13] It should be noted that it is even possible to combine the two approaches, that is, to account for differences in some of the model parameters using continuous random effects and in some other parameters using the discrete approach.

There are too many possible model specifications of a multilevel LC model to discuss them all in detail. What is clear is that the model described by Vermunt[4,16] in which the $P(x_{ji}=k|G=j)$ parameters are allowed to differ across groups is the most important one. This model assumes that observations within groups are correlated because group members tend to belong to the same LC. The model is obtained by defining a random effects multinomial logistic regression model for class membership, using either a continuous or a discrete specification of the random effects. Vermunt[4,16] proposed

a model in which the parameters defining the class-specific densities for the response variables do not vary across groups ($\boldsymbol{\vartheta}_{jtk} = \boldsymbol{\vartheta}_{tk}$) and in which

$$\log \frac{P(x_{ji} = k|G = j)}{P(x_{ji} = 1|G = j)} = \gamma_k + \tau_k \cdot u_j$$

for $2 \leq k \leq K$, where u_j is a normally distributed random effect or latent variable with a mean equal to 0 and a variance equal to 1 and τ_k are factor loadings. Note that this specification yields a random-effects multinomial logit model for class membership that is similar to the restricted random-effects multinomial logit model – containing one instead of $K - 1$ random effects – proposed by Hedeker[31] for observed categorical responses.

The current article focuses on the discrete approach with group-level latent classes. The main extension compared to the work by Vermunt[16] is that it is not only applied to categorical response variables as in the traditional LC model but also to continuous variables and counts. The basic idea of the discrete approach is that groups belong to one of L group-level LCs or mixture components, where w_j denotes the class membership of group j and ℓ is used to refer to a particular group-level LC, $1 \leq w_j = \ell \leq L$. The multilevel LC model can now be defined as follows:

$$
\begin{aligned}
f(y_{ji}|w_j = \ell) &= f(y_{ji}; \boldsymbol{\vartheta}_\ell) \\
&= \sum_{k=1}^{K} P(x_{ji} = k|w_j = \ell) \prod_{t=1}^{T} f(y_{jit}|x_{ji} = k, w_j = \ell) \\
&= \sum_{k=1}^{K} P(x_{ji} = k|w_j = \ell) \prod_{t=1}^{T} f(y_{jit}; \boldsymbol{\vartheta}_{tk\ell})
\end{aligned}
\tag{3}
$$

As can be seen, LCs of groups may differ in the probability that their members belong to LC k and in the parameters defining the densities for the response variables.

The element in the model specification that is still missing is the connection between subjects belonging to the same group, that is,

$$f(y_j) = \sum_{\ell=1}^{L} P(w_j = \ell) \prod_{i=1}^{n_j} f(y_{ji}; \boldsymbol{\vartheta}_\ell) \tag{4}$$

As can be seen, the density for the full response vector of group j, $f(y_j)$, is obtained with the additional assumption that its n_j members' responses are independent of one another conditional on group class membership and by subsequently marginalizing over the LCs for groups. In the case of continuous random effects, the summation over the group-level mixture components is replaced by an integration over the random effects, and it is assumed that observations within groups are independent of one another given these group-level random effects.[12]

It should be noted that the multilevel LC model is actually a model for three-level data sets, that is, for multiple level-1 observations (responses) within level-2 units (subjects) and multiple level-2 units within level-3 units (groups). This means that the model can also be used to deal with other types of three-level data than multiple responses nested within subjects nested within groups. One of the examples discussed in the next section deals with multiple prescriptions of doctors working in different hospitals. Another example deals with longitudinal data in which level-2 units represent time points and level-3 units represent subjects. The fact that this is a three-level random-effects model shows that it could easily be generalized to a larger number of levels using the same logic as was used to go from two to three levels.

Vermunt[13,16] focussed on two restricted special cases of the general model described in Equations (3) and (4), namely:

- a model in which subject-level class membership probabilities differ across group-level classes but in which the parameters defining the class-specific conditional distributions for the response variables do not vary across group-level classes ($\boldsymbol{\vartheta}_{tk\ell} = \boldsymbol{\vartheta}_{tk}$)
- a model in which a GLM regression model is defined for the response variables y_{jit} with additive effects of x_{ji} and w_j – for example, $\eta_{jit} = \beta_{0t} + \beta_{1k} + \beta_{2\ell}$, where η_{jit} is the linear predictor of the GLM concerned for y_{jit} – but in which subject-level class membership probabilities do not differ across group-level classes [$P(x_{ji} = k | w_j = \ell) = P(x_{ji} = k)$].

In the former specification, the LCs at the lower-level capture all the association between the responses within lower-level units, whereas the higher-level classes capture the association between subjects within groups. This is, in fact, a two-level variant of a standard LC model. The latter specification is more similar to a standard three-level regression model: the higher-level classes capture the common variation of all responses within a group and the lower-level classes the common residual variation within subjects, yielding the well known variance decomposition between levels of the hierarchy. This difference shows that the former specification is more interesting if the main focus is on identifying meaningful lower-level classes taking into account the multilevel data structure, whereas the latter specification is more meaningful when the finite mixture model is used as a nonparametric three-level random effects regression model. In the applications discussed below, the author would focus on the former specification. For more information on the latter specification, see Vermunt.[13]

As was shown by Vermunt,[4,13,16] ML estimates of the model parameters of the multilevel LC model can be obtained by means of the expectation-maximization (EM) algorithm.[32] This, however, requires a special implementation of the expectation (E) step of the EM algorithm in which the relevant posteriors are computed in an efficient way by making use of the conditional independence assumptions implied by the multilevel LC model. This version of the EM algorithm which is described in the appendix, as well as a Newton–Raphson algorithm with analytic first-order derivatives, and numerical second-order derivatives, is implemented in the Latent GOLD software package.[23] Some versions of the multilevel LC model can be estimated with the GLLAMM software[30] and with Mplus.[33] More specifically, GLLAMM can estimate LC models with discrete

random effects in the model for the response variable, and Mplus can estimate LC models with continuous group-level random effects in the model for the subject-level classes. As an alternative to ML estimation, one may adopt a Bayesian Markov chain Monte Carlo (MCMC) estimation approach, for example, using the WinBugs software,[34] in which it is not difficult to program the multilevel LC model (see the appendix). Bayesian estimation of mixture models is, however, complicated by the fact that the arbitrary ordering of the classes may change during the MCMC chain, which is usually referred to as the label-switching problem.[35,36] This problem is even harder to deal with when there are multiple mixture distributions like in the multilevel LC model described in this article.

4   Examples

In this section, three rather different types of applications of the multilevel LC model are presented. The first application uses a typical three-level data set: prescriptions of antibiotics are nested within doctors which are themselves nested within hospitals. The second application is a more standard cluster analysis application in which multiple continuous responses (indicators) are used to classify subjects: there are six measures of intelligence of children which are nested within families. In the last application, lower-level LCs capture the time-specific differences and higher-level classes the stable differences between individuals in a count response variable that is measured at multiple time points. For parameter estimation, the Latent GOLD software[23] is used (see the appendix for model syntax files).

4.1   Prescription of antibiotics:

The data set used for the first example is described in detail by Yan and by Rabe-Hesketh and Skrondal.[30,37] The outcome variable 'abuse' is the classification of 855 antibiotics prescriptions of 134 doctors in two Chinese counties as 1) correct use; 2) abuse of one antibiotic; 3) abuse of several antibiotics in the treatment of children with acute respiratory tract infection. The purpose of the LC analysis is to identify clusters of doctors with similar prescription behaviour. Doctors belong to 36 hospitals, which are the higher-level units. In the multilevel LC model, this nesting is taken into account by allowing hospitals to belong to higher-level classes that differ in the proportion of doctors with a particular prescription behaviour. In a second phase of the analysis, the effect of the doctor's education and the dummy for hospital's participating in a WHO programme that was initiated as a response to the misuse of antibiotics, were investigated.

Table 2 presents the log-likelihood (LL) value, the Bayesian information criterion (BIC), and the number of parameters (Npar) for the estimated models. In the computation of BIC, I use the number of subjects (doctors) as the sample size; that is, $BIC = -2LL + Npar \log N$, with $N = \sum_{j=1}^{J} n_j$. As can be seen, the model with two hospital-level classes ($L = 2$) and three doctor-level classes ($K = 3$) performs best according to the BIC. The same model would be selected using the less well-known AIC(3) criterion that is obtained by replacing the constant $\log N$ in the BIC formula by the

Table 2 Fit measures for the models estimated with the antibiotics data set

L	K	LL	BIC	Npar
1	1	−914.3	1838.3	2
1	2	−872.4	1769.3	5
1	3	−865.1	1769.3	8
1	4	−859.4	1772.7	11
2	2	−859.6	1753.5	7
2	3	−846.7	1747.3	11
2	4	−841.6	1756.6	15
3	2	−859.6	1763.3	9
3	3	−844.3	1757.1	14
3	4	−837.3	1767.6	19

value 3 and that has recently been suggested as possible model selection measure in the context of mixture modelling.[38]

Table 3 presents the parameter estimates for this best fitting model. The lower part of this table reports the abuse probabilities for each of the doctor-level classes. Compared with the other two classes, doctors in class 1 have the highest probability of no abuse, and in the case of abuse they are more likely to be abusing 'only' one antibiotic. Class 3 has the highest probability of abuse of several antibiotics, whereas doctors in class 2 have the highest probability of abusing one antibiotic.

The upper part of Table 3 indicates that 50% of the hospitals belong to higher-level LC 1 and the other 50% to class 2. These two hospital-level classes have very different distributions of doctors among doctor-level classes. Almost all doctors belonging to the first hospital class belong to the low abuse doctor class, whereas almost all doctors in the other hospital class are in the middle and high abuse classes. Overall, we can conclude that there are large differences in the abuse of antibiotics between hospitals and doctors.

Then two covariates were introduced in the model. It was assumed that doctor's education measured in six ordinal categories affects the probability of being in one of the doctor-level classes and that the dichotomous hospital-level variable for being in the WHO programme affects the probability of belonging to one of the hospital-level classes. This gives a model with LL = −830.4 and Npar = 14, which indicates that at least one of the effects is significant. Two separate Wald tests yield values equal to

Table 3 Hospital-class [$P(w_j = \ell)$], doctor-class conditional on hospital class [$P(x_{ji} = k | w_j = \ell)$] and doctor-class-specific response probabilities (π_{tkm}) obtained with the model with $L = 2$ and $K = 3$ estimated with the antibiotics data set

	$P(w_j = \ell)$	$k = 1$ (low)	$k = 2$ (middle)	$k = 3$ (high)
$\ell = 1$ (low)	0.50	0.91	0.01	0.08
$\ell = 2$ (high)	0.50	0.06	0.33	0.61
Abuse = 1		0.39	0.11	0.29
Abuse = 2		0.56	0.64	0.21
Abuse = 3		0.05	0.25	0.50

16.33 (d$f = 2$; $P < .001$) and 7.94 (d$f = 1$; $P = 0.005$) for education and WHO pro-
gramme, respectively, showing that both effects are significant. The definition of both
the higher- and lower-level classes changed somewhat, but the substantive conclusion
is that higher educated doctors are (much) more likely to be in the low-abuse class,
whereas lower educated doctors are more likely to be in the high-abuse class. WHO
hospitals are much more likely to be in hospital-level class containing a large pro-
portion doctors in the low-abuse class, which means that the programme seems to be
effective.

The same model was also estimated with three doctor and two hospital classes (with-
out covariates) with Winbugs using Dirichlet priors with parameters equal to 1 for the
multilevel LC model probabilities (see model setup in the appendix). The obtained
Bayesian parameter estimates and standard errors were very close to the ML esti-
mates. To deal with the label-switching problem, the parameters were estimated using
500 samples from a part of the MCMC chain in which label switching did not seem
to occur. Note that the label-switching problem also complicates convergence checks
based on running multiple chains since the classes may have a different order in each
chains.

4.2 **Intelligence test**

The data for this example was collected by Van Peet[39] and used by Hox[29] to illustrate
multilevel FA. Six continuous measures supposed to be connected to intelligence – word
list, cards, figures, progressive matrices, name animals and name occupations – are
available for 269 children belonging to 49 families. For 82 children, there is partially
missing information, but these observations can be retained using standard ML method-
ology with missing data. The six intelligence measures are used to cluster the children
into intelligence classes and it is investigated whether (classes of) families differ in the dis-
tribution of children over these 'intelligence' clusters. The differences between families
can be attributed to genetic and/or common environment effects.

Preliminary analysis showed that simple univariate normal within-class distributions
can be assumed for the six response variables with equal residual variances across classes.
More specifically, inspection or pairwise residuals showed that there is no need to allow
for within-class correlations across responses, and comparison of models with equal
and unequal variances showed that it is correct to assume that residual variances are
homogeneous across classes.

Table 4 provides the fit measures for the estimated multilevel LC models, where the
specification was that family-level classes affect child-level classes and child-level classes
affect the responses. As can be seen, a model with four child-level classes and three
family-level classes performed best according to the BIC. As in the previous application
the AIC(3) criterion selects the same model as the best one.

Table 5 provides the parameters obtained with this model (the within-class variances
for the six indicators which do not appear in this table equal 28.3, 30.6, 26.1, 29.9, 26.1
and 26.5, respectively). The lower part of this table shows that the class-specific means
of the six intelligence indicators are nicely ordered across child-level classes 1–3. These
can therefore be labelled as high middle and low. Children in class 4 show a somewhat
mixed pattern: they perform better than the middle class on cards and figures, better

Table 4 Fit measures for the models estimated with the intelligence data set

L	K	LL	BIC	Npar
1	1	−4237.7	8542.5	12
1	2	−4148.8	8403.9	19
1	3	−4127.2	8399.8	26
1	4	−4113.2	8411.0	33
1	5	−4104.2	8432.2	40
2	2	−4130.2	8377.9	21
2	3	−4108.4	8379.0	29
2	4	−4086.8	8380.6	37
2	5	−4075.2	8402.2	45
3	2	−4129.7	8388.0	23
3	3	−4097.6	8374.2	32
3	4	−4072.2	8373.7	41
3	5	−4060.3	8400.3	50
4	2	−4129.6	8399.1	25
4	3	−4096.0	8387.9	35
4	4	−4069.7	8391.3	45
4	5	−4052.2	8412.1	55

than the low class on word list and matrices and worse than the low class on animals and occupations.

The upper part of Table 5 shows that in the family-level class 3, almost all children belong to the the mixed child-level class. Children from families belonging to family-level class 1 are more likely to be in the high intelligence class and children from family-level class 2 are more often in the middle and low intelligence classes. These results show that there is a very strong family effect on the performance of children on these six intelligence subtests.

One child-level covariate (gender) was available and was used it as a predictor of the child's likelihood of being in one of the intelligence clusters. Its effect turned out to be significant (LL = −4063.9, Npar = 44). The logit coefficients indicate that boys are more likely than girls to be in class 1 and that girls are more likely to be in class 4 than boys.

Table 5 Family-class probabilities [$P(w_j = \ell)$], child-class probabilities conditional on family class [$P(x_{ji} = k | w_j = \ell)$] and the child-class-specific normal means (μ_{tk}) obtained with the model with $L = 3$ and $K = 4$ estimated with the intelligence data set

	$P(w_j = \ell)$	$k = 1$ (high)	$k = 2$ (middle)	$k = 3$ (low)	$k = 4$ (mixed)
$\ell = 1$	0.56	0.74	0.25	0.00	0.02
$\ell = 2$	0.32	0.17	0.59	0.24	0.01
$\ell = 3$	0.12	0.07	0.01	0.03	0.90
Wordlist		31.9	29.4	25.4	26.2
Cards		36.0	30.0	22.5	34.0
Figures		29.0	26.1	22.2	27.2
Matrices		34.0	29.9	26.8	28.8
Animals		30.5	28.4	24.1	21.5
Occupations		29.1	28.5	25.5	23.0

4.3 Repeated measures epilepsy data

For the third empirical example, the longitudinal epilepsy data set from Thall and Vail[40] coming from a randomized trial that was setup to determine the effect of a new treatment of epilepsy was used. The outcome variable is the number of epileptic seizures during the two weeks before each of the four clinic visits. The number of patients is 59. The mixture model is used as a tool for dealing with overdispersed count data. More specifically, at each of the occasions, patients can be in one of K LCs that differ with respect to the mean of the Poisson distribution. The measurements at the four occasions are connected by a discrete latent variable at the subject level. The substantive question of interest is, of course, whether there is evidence that the new treatment is effective.

Table 6 presents the BIC values for the various models estimated with the epilepsy data set. We again use the specification in which higher-level classes affect lower-level classes and lower-level classes affect the responses. According to BIC, $L = 2$ and $K = 5$ is the best model for this data set, which means that there are five time-specific latent states and two permanent latent states. In this application, BIC and AIC(3) do not point at the same model: according to AIC(3), the number of permanent classes should be three instead of two.

The parameter estimates for the model with $L = 2$ and $K = 5$ reported in Table 7 show that there are huge differences in the average number of seizures among the five time-specific LCs: the Poisson means range from 0.3 to 75.3. The upper part of the table shows that the two time-constant classes consisting of 69 and 31 patients show clear difference in their likelihood of being in one of the time-specific classes: the first higher-level class is much more likely to belong to the two small number of seizure classes and the second to the three large number of seizure classes. The model with $L = 3$ and $K = 5$ yields very similar results: the definition of the time-specific classes does not change, whereas the three permanent classes can be ordered from low to high.

As a next step, it was investigated whether treatment has an effect on the logit of belonging to the healthier higher-level class and whether the distribution of the

Table 6 Fit measures for the models estimated with the epilepsy data set

L	K	LL	BIC	Npar
1	1	−1641.9	3289.2	1
1	2	−945.2	1906.8	3
1	3	−771.2	1569.6	5
1	4	−741.2	1520.5	7
1	5	−725.7	1500.6	9
1	6	−723.3	1506.7	11
2	2	−914.0	1855.2	5
2	3	−724.1	1491.8	8
2	4	−685.2	1430.4	11
2	5	−670.8	1418.1	14
2	6	−667.9	1428.6	17
3	2	−911.6	1861.4	7
3	3	−716.3	1492.7	11
3	4	−671.3	1424.6	15
3	5	−658.8	1421.4	19
3	6	−654.0	1433.6	23

Table 7 Time-constant class probabilities [$P(w_j = \ell)$], time-varying class probabilities conditional on time-constant class [$P(x_{ji} = k|w_j = \ell)$] and time-varying class Poisson means (θ_k) obtained with the model with $L = 2$ and $K = 5$ estimated with the epilepsy data set

	$P(w_j = \ell)$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
$\ell = 1$ (low)	0.69	0.15	0.81	0.03	0.00	0.00
$\ell = 2$ (high)	0.31	0.03	0.02	0.56	0.32	0.07
Seizures		0.3	3.8	10.7	23.7	75.3

lower-level classes changed over time. Neither treatment nor time turned out to have a significant effect: LL = −670.6, Npar = 15 and L = −667.9, Npar = 26.

5 Discussion

This article presented an extension of LC and finite mixture models for dealing with multilevel data sets. The variant that received most attention is a model in which higher-level units are clustered based on the (prior) class membership probabilities of their members. Although, the approach was originally proposed by Vermunt[16] for LC models with categorical response variables, it was shown that the methodology can also be used in context of finite mixture models for continuous response variables and counts.

The modelling approach was illustrated with three applications, a typical three-level application, a clustering application and a repeated measures clinical trial. In each of these three examples, it turned out that there was a need to take into account the hierarchical nature of the data set when defining an LC model. Also from a substantive point of view, the multilevel LC model turned out to have a clear added value compared with the standard LC model. In the first example, classes of hospital which contain doctors with very different prescription behaviours were found. Hospitals participating in a WHO programme performed much better than hospitals that did not participate in this programme. The second example showed a very strong family effect on the performance of children on six intelligence tests, with one family cluster containing children with a mixed intelligence level. The last example illustrated the modelling of longitudinal count data using the multilevel LC model. From a substantive point of view, the results were, however, somewhat atypical: there was no evidence at all for a treatment effect, and there was no evidence that class membership probabilities changed during the observation period.

As far as the discrete versus continuous specification for the latent variables at levels 2 and 3 is concerned, it can be observed that especially in the analysis of the intelligence data set the discrete approach turned out to be extremely valuable from a substantive point of view. Interesting and easy to explain patterns were encountered that would never have been detected using a model with continuous latent variables. In the other two applications, the discrete latent variables turned out to be either ordered or dichotomous, which implies that similar substantive conclusions would have been reached using models with continuous latent variables. The only advantage of the discrete specification was that it does not require making strong assumptions about the distributions of the latent variables.

Various interesting extensions of the multilevel LC model are possible. The most straightforward one would be an extension to data sets with additional levels in the hierarchy. There are no practical or theoretical obstacles for such an extension: as shown by Vermunt,[13] the upward–downward variant of the EM algorithm can easily be extended to any number of levels.

Another extension to be mentioned is the possibility of using continuous latent variables in addition to classes at the lower level. This yields mixed variants of well-known latent variable models such as item response models and FA. Within a multilevel context with discrete random effects at the higher level, one obtains a very interesting latent variable model with mixture distributions at both the subject and the group level, each of which would capture a different type of unobserved heterogeneity.

Appendix

Implementation of the *E* step of the EM algorithm

By treating the w_j and x_{ji} variables as missing or unobserved, parameter estimation by ML can be solved by means of the EM algorithm.[32] Because of the extremely high dimensionality of the missing data problem, in the implementation of the *E* step similar trick as in the Baum–Welch algorithm for hidden Markov modelling was used.[41] It should be noted that the model defined in Equations (3) and (4) contains $1 + n_j$ unobserved variables with a total of $L \cdot K^{n_j}$ categories. This implies, for example, that with $n_j = 8$ and $K = L = 4$, a model with 262 144 entries in the joint distribution of the discrete variables with missing values is dealt with. It will be clear that this can not be solved with a standard EM algorithm. In the graphical or Bayesian belief network modelling field, the multilevel LC model would be recognized as a single-connected network or polytree, for which marginal probabilities for subsets of variables can be obtained by propagation algorithms.[42] Both the forward–backward algorithm for hidden Markov models and the upward–downward algorithm discussed below are propagation algorithms.

Rather than repeating all the well-known details on the EM algorithm for the estimation of finite mixture models which can be found in, for example, McLachlan and Peel,[9] the specific aspects associated with the estimation of the multilevel LC model described in Equations (3) and (4) will be concentrated on. The expected value of the complete data LL function for this model has the following form:

$$
\begin{aligned}
E[\log L_C(\boldsymbol{\vartheta})] = & \sum_{j=1}^{J} \sum_{\ell=1}^{L} \widehat{w}_{j\ell} \log P(w_j = \ell) \\
& + \sum_{j=1}^{J} \sum_{\ell=1}^{L} \sum_{i=1}^{n_j} \sum_{k=1}^{K} \widehat{w}_{j\ell} \widehat{x}_{jik|\ell} \log P(x_{ji} = k | w_j = \ell) \\
& + \sum_{j=1}^{J} \sum_{\ell=1}^{L} \sum_{i=1}^{n_j} \sum_{k=1}^{K} \widehat{w}_{j\ell} \widehat{x}_{jik|\ell} \log f(\boldsymbol{y}_{ji} | \boldsymbol{\vartheta}_{k\ell})
\end{aligned}
$$

where $\widehat{w}_{j\ell} = P(w_j = \ell | \mathbf{y}_j; \boldsymbol{\vartheta})$ and $\widehat{x}_{jik|\ell} = P(x_{ji} = k | w_j = \ell, \mathbf{y}_j; \boldsymbol{\vartheta})$ denote the posterior probability that group j belongs to class ℓ and that subject i belongs to class k given that the group belongs to class ℓ, conditional on the observed data and the current parameter estimates. Calculation of the expected value of the complete data LL – which is the E step of the EM algorithm – thus involves computing these two sets of posterior probabilities. Note that $\widehat{w}_{j\ell} \cdot \widehat{x}_{jik|\ell} = P(w_j = \ell, x_{ji} = k | \mathbf{y}_j; \boldsymbol{\vartheta})$.

Crucial in the implementation of the E step of the EM algorithm is that one can make use of the fact that lower-level observations are independent of one another given the higher-level (group) class memberships. More specifically, we make use of the fact that

$$\widehat{x}_{jik|\ell} = P(x_{ji} = k | w_j = \ell, \mathbf{y}_j; \boldsymbol{\vartheta}) = P(x_{ji} = k | w_j = \ell, \mathbf{y}_{ji}; \boldsymbol{\vartheta})$$

that is, that given class membership of the group (w_j), class membership of a subject (x_{ji}) is independent of the observed data of other subjects in the group concerned.

In order to simplify the formulas for $\widehat{w}_{j\ell}$ and $\widehat{x}_{jik|\ell}$, let

$$h_{jik|\ell} = P(x_{ji} = k | w_j = \ell) f(\mathbf{y}_{ji} | x_{ji} = k, w_j = \ell)$$
$$= P(x_{ji} = k | w_j = \ell) f(\mathbf{y}_{ji} | \boldsymbol{\vartheta}_{k\ell})$$

and

$$g_{ji|\ell} = \sum_{k=1}^{K} h_{jik|\ell}$$

The relevant terms are obtained as follows:

$$\widehat{w}_{j\ell} = \frac{P(w_j = \ell) \prod_{i=1}^{n_j} g_{ji|\ell}}{\sum_{\ell=1}^{L} P(w_j = \ell) \prod_{i=1}^{n_j} g_{ji|\ell}}$$

$$\widehat{x}_{jik|\ell} = \frac{h_{jik|\ell}}{g_{ji|\ell}}$$

As can be seen, for each group j, we first compute $h_{jik|\ell}$ for each ℓ, i and k combination and collapse these over k to obtain $g_{ji|\ell}$, which amounts to marginalizing over the lower-level discrete latent variables. Combining the $g_{ji|\ell}$ for all i gives the posterior for the higher-level LC variable. Analogous to the forward–backward algorithm, Vermunt[16] refers to these steps as the upward steps. The downward step involves the computation of $\widehat{x}_{jik|\ell}$ and the bivariate joint posterior of w_j and x_{ji}, the (product) term that enters in the expected complete data LL; that is,

$$P(w_j = \ell, x_{ji} = k | \mathbf{y}_j; \boldsymbol{\vartheta}) = \widehat{w}_{j\ell} \cdot \widehat{x}_{jik|\ell}$$

In M step of the EM algorithm, standard complete data methods can be used for finding estimates of the unknown model parameters that maximize (or increase) the expected complete data LL.

Latent GOLD 4.5 syntax for examples

The models discussed in this article can be defined using the Latent GOLD point and click graphical user interface. In the newest version Latent GOLD 4.5,[43] it is also possible to define models using a syntax language, which makes the model structure much more transparent. More specifically, the model definition consists of a series of regression equations for the latent and the response variables, as well as of the specification of the settings for the (residual) variances and covariances. Before defining the regression equations, one has to specify the technical and output options and the id variables, as well as the names and scale types of the latent, dependent and independent variables that play a role in the model. Scale types are denoted with keywords such as nominal, ordinal, continuous, numeric, poisson and binomial.

In the first example, a long file with a single nominal dependent variable Abuse, was used. The records of a doctor are linked by a case-id variable and the records of a hospital by a group-id. `HospitalClass` and `DoctorClass` are discrete latent variables at the hospital and doctor level. The model definition consists of three regression equations:

```
HospitalClass <- 1;
DoctorClass <- 1 + HospitalClass;
Abuse <- 1 + DoctorClass;
```

Here, '1' is used to define the constant in the regression equation concerned. The model with the predictors `WHO` and `Education` is defined as follows

```
HospitalClass <- 1 + WHO;
DoctorClass <- 1 + HospitalClass + Education;
Abuse <- 1 + DoctorClass;
```

As was indicated in the text, `HospitalClass` may also be used as a predictor in the model for the response variable Abuse rather than in the model for the `DoctorClass`:

```
HospitalClass <- 1;
DoctorClass <- 1;
Abuse <- 1 + DoctorClass + HospitalClass;
```

This yields a more standard 3-level regression model with discrete level-2 and level-3 random effects in the model for the response variable.

In the second example, a standard rectangular data file containing the six continuous response variables in separate columns was used. A group-id connects the records of children belonging to the same family. The multilevel LC model is specified as follows:

```
FamilyClass <- 1;
ChildClass <- 1 + FamilyClass;
WordList <- 1 + ChildClass;
Cards <- 1 + ChildClass;
Figures <- 1 + ChildClass;
Matrices <- 1 + ChildClass;
Animals <- 1 + ChildClass;
Occupations <- 1 + ChildClass;
WordList;
Cards;
Figures;
```

```
Matrices;
Animals;
Occupations;
```

Note that the last six lines define the structure for the residual variances. Although in this example a specification with homoscedastic residuals was used, it is straightforward to define models with heteroscedastic errors. For example, if the error variances are allowed to depend on `FamilyClass`, then 'Wordlist;' is replaced by 'WordList | FamilyClass;', etc. Similarly, it could be indicated that residual variances differ across child-level LCs.

The setup for the third example is similar to the one for the first example, with the only difference that no id variable is needed at the lower level since there is only one response per time point ($T = 1$).

Winbugs model definition

For the prescription of antibiotics example, the model was also estimated using the MCMC approach implemented in Winbugs 1.4. The model definition for a model with L and K classes is as follows:

```
for (jit in 1:855) {abuse[jit] ~ dcat(piy[x[docid[jit]],])}
for (ji in 1:134) {x[ji] ~ dcat(pix[w[hospid[ji]],])}
for (j in 1:36) {w[j] ~ dcat(piw[])}
for (k in 1:K) {piy[k,1:3] ~ ddirch(prior[1:3])}
for (l in 1:L) {pix[l,1:K] ~ ddirch(prior[1:K])}
piw[1:L] ~ ddirch(prior[1:L])
```

In the data file, one defines the number of doctor- and hospital-level LCs (L and K), the value of the response variable (abuse) and the doctor-id (docid) for each prescription, the hospital-id (hosid) for each doctor and the hyper-parameters of the Dirichlet priors (each entry was set equal to 1, which amounts to using diffuse priors). In addition, one may wish to specify initial values for the model parameters (piy, pix and piw).

References

1 Rindskopf D, Rindskopf W. The value of latent class analysis in medical diagnosis. *Statistics in Medicine* 1986; **5**: 21–7.

2 Uebersax JS, Grove WM. Latent class analysis of diagnostic agreement. *Statistics in Medicine* 1990; **9**: 559–72.

3 Bandeen-Roche KJ, Miglioretti DL, Zeger SL, Rathouz PJ. Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association* 1997; **92**: 1375–86.

4 Vermunt JK. Mixed-effects logistic regression models for indirectly observed outcome variables. *Multivariate Behavioral Research* 2005; **40**: 281–301.

5 Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 1974; **61**: 215–31.

6 Lazarsfeld PF, Henry NW. *Latent structure analysis*. Houghton Mifflin, 1968.

7 Magidson J, Vermunt JK. Latent class factor and cluster models, bi-plots and related graphical displays, *Sociological Methodology* 2001; **31**: 223–64.

8 Everitt BS (1988). A finite mixture model for the clustering of mixed-mode data. *Statistics and Probability Letters* 1988; **6**: 305–09.

9 McLachlan GJ, Peel D. *Finite mixture models*. John Wiley & Sons, 2000.

10 Vermunt JK, Magidson J. Latent class cluster analysis. In Hagenaars JA, McCutcheon A eds. *Applied Latent Class Analysis*. Cambridge University Press, 2002, 89–106.

11 Aitkin M. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* 1999; **55**: 218–34.

12 Skrondal A, Rabe-Hesketh S. *Generalized latent variable modeling: multilevel, longitudinal and structural equation models*. Chapman & Hall/CRC, 2004.

13 Vermunt JK. An EM algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistical Neerlandica* 2004; **58**: 220–33.

14 Goldstein H. *Multilevel statistical models*, third edition. Arnold, 2003.

15 Snijders TAB, Bosker RJ. *Multilevel analysis*. Sage Publications, 1999.

16 Vermunt JK. Multilevel latent class models. *Sociological Methodology* 2003; **33**: 213–39.

17 Fox J-P, Glas CAW. Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* 2001; **66**: 269–86.

18 Goldstein H, Browne W. Multilevel factor analysis modelling using Markov chain Monte Carlo estimation. In Marcoulides GA, Moustaki I eds. *Latent variable and latent structure models*. Lawrence Erlbaum Associates, 2002: 225–43.

19 Longford N, Muthén B. Factor analysis for clustered observations. *Psychometrika* 1992; **57**: 581–97.

20 Rabe-Hesketh S, Skrondal A, Pickles A. Generalized multilevel structural equation modelling. *Psychometrika* 2004; **69**: 167–90.

21 Raudenbush SW, Johnson C, Sampson RJ. A multivariate, multilevel Rasch model with application to self-reported criminal behavior. *Sociological Methodology* 2003; **33**: 169–211.

22 Vermunt JK. Multilevel mixture item response theory models: an application in education testing. Proceedings of the 56th session of the International Statistical Institute. Lisbon, Portugal, 22–28 August, 2007.

23 Vermunt JK, Magidson J. *Latent GOLD 4.0 user's guide*. Statistical Innovations Inc., 2005.

24 Palardy G, Vermunt JK. Multilevel growth mixture models for classifying group-level observations. Submitted for publication, 2007 http://spitswww.uvt.nl/~vermunt/palardy 2007.pdf.

25 Hagenaars JA. Latent structure models with direct effects between indicators: local dependence models. *Sociological Methods and Research* 1988; **16**: 379–405.

26 Dayton CM, Macready GB. Concomitant-variable latent-class models. *Journal of the American Statistical Association* 1988; **83**: 173–78.

27 Hadgu A, Qu Y. A biomedical application of latent class models with random effects. *Applied Statistics* 1998; **47**: 603–16.

28 Muthén B. Latent variable mixture modeling. In Marcoulides GA, Schumacker RE eds. *New developments and techniques in structural equation modeling*. Lawrence Erlbaum Associates, 2001: 1–33.

29 Hox J. *Multilevel analysis: techniques and applications*. Lawrence Erlbaum, 2002.

30 Rabe-Hesketh S, Skrondal A. *Multilevel and longitudinal modeling using Stata*. Stata Press, 2005.

31 Hedeker D. A mixed-effects multinomial logistic regression model. *Statistics in Medicine* 2003; **22**: 1433–46.

32 Dempster AP, Laird NM, and Rubin DB. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*. 1997; **39**: 1–38.

33 Muthén LK, Muthén BO. *Mplus user's guide*, fourth edition. Muthén & Muthén, 1998–2006.

34 Spiegelhalter D, Thomas A, Best N, Lunn D. *Winbugs user manual version 1.4*. MRC Biostatistics Unit at Institute of Public Health at Cambridge University and Department of Epidemiology & Public Health at Imperial College School of Medicine, 2003.

35 Dias JG, Wedel M. An empirical comparison of EM, SEM and MCMC performance for problematic Gaussian mixture likelihoods. *Statistics and Computing* 2004; **14**: 323–32.

36 Stephens M. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society B* 2000; **62**: 795–809.

37 Yan M. Multinomial regression. In Leyland AH, Goldstein H eds. *Multilevel modelling in health statistics*. Wiley, 2001: 107–23.

38 Andrews RL, Currim IS. A Comparison of segment retention criteria for finite mixture logit models, *Journal of Marketing Research* 2003; **40**: 235–43.

39   Van Peet AAJ. *De potentieeltheorie van intelligentie (The potentiality theory of intelligence)*. University of Amsterdam, PhD thesis, 1992.

40   Thall PF, Vail SC. Some covariance structure model for longitudinal count data with overdispersion. *Biometrics* 1990; **46**: 657–71.

41   Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics* 1970; **41**: 164–71.

42   Pearl J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

43   Vermunt JK, Magidson J. *LG-syntax user's guide: manual for Latent GOLD 4.5 syntax module*. Statistical Innovations Inc., 2007.