Chapter 1
Mixture Model Clustering with Covariates Using Adjusted Three-step Approaches

Dereje W. Gudicha and Jeroen K. Vermunt

Abstract When using mixture models, researchers may investigate the associations between cluster membership and covariates by introducing these variables in a (logistic) regression model for the prior class membership probabilities. However, a very popular alternative among applied researchers is a three-step approach in which after estimating the mixture model (step 1) and assigning subjects to clusters (step 2), the cluster assignments are regressed on covariates (step 3). For mixture models for categorical responses, Bolck et al. (2004) and Vermunt (2010) showed this approach may severely downward bias covariate effects, and moreover showed how to adjust for this bias. This paper generalizes their corrections methods to be applicable also with mixture models for continuous responses, where the main complicating factor is that a complex multidimensional integral needs to be solved to obtain the classification errors needed for the corrections. We propose approximating this integral by a summation over the empirical distribution of the response variables. The simulation study showed that the approaches work well, except for the combination of very badly separated components and a small sample size.

1.1 Introduction

Most applied researchers using mixture models not only aim at finding a meaningful set of clusters, but also wish to investigate which factors are associated with the cluster membership of subjects. This profiling of clusters (or latent classes) as a function of external variables (covariates) can either be achieved using a one-step

Dereje W. Gudicha
Tilburg University, PO Box 50193, 5000 LE Tilburg, The Netherlands, e-mail: d.w.gudicha@uvt.nl

Jeroen K. Vermunt
Tilburg University, PO Box 50193, 5000 LE Tilburg, The Netherlands, e-mail: j.k.vermunt@uvt.nl

approach or a three-step approach (Bolck et al., 2004). In the one-step approach, the mixture model is expanded by including the relevant covariates in a regression model for the prior class membership probabilities (Bandeen-Roche et al., 1997; Dayton and Macready, 1988). The parameters defining the mixture components – the cluster specific means and (co)variances – and the covariate effects on the cluster membership are estimated simultaneously. Alternatively, in the much more popular three-step approach, the analysis is done in a stepwise manner. First, a standard mixture model clustering is performed without covariates; second, the class membership is predicted, typically using the Bayes modal rule; third, the association between external variables and the predicted class membership is assessed, for example, via a logistic regression analysis. Bolck et al. (2004) and Vermunt (2010) showed for latent class models with categorical responses that this three-step approach may yield severely downward biased estimates for the covariate effects. These authors also showed how to adjust for this bias in step three by using information on the classification errors introduced in step two. Bolck et al. proposed a weighted analysis with the inverse of the classification errors as weights whereas Vermunt proposed a maximum likelihood method that takes the classification errors into account.

While the use of mixture models with continuous response variables is very common, it is not immediately clear how the adjusted three-step methods should be implemented when the response variables used in the mixture model are not categorical but continuous. The aim of the current paper is to come up with such a generalization. The main complicating factor is that the computation of the classification error matrix needed in step three requires solving a complex multidimensional integral. We propose using Monte Carlo integration for this purpose, which if the model holds can be replaced by a summation over the observed data points. The performance of this approach is investigated in a simulation study.

The remainder of this paper is organized as follows. First, the mixture model of normal distributions is introduced and the estimation of the class memberships and the quantification of the classification errors is discussed. Subsequently, the relevant one- and three-step approaches for investigating the association between external variables and class membership are presented. These approaches are evaluated in a simulation study. The paper ends with conclusions and practical recommendations.

1.2 Mixture Modeling and Classification

1.2.1 Mixture Models

The first step of the three-step approach involves estimating the parameters of a mixture model without covariates (i.e., the class proportions and the cluster specific means and (co)variances). Suppose that we have information on p response variables and that the interest lies in clustering of n observations into k exhaustive and mutually exclusive homogeneous subgroups (latent classes). Let T be an

unobservable random variable containing the labels of the k subpopulations with realizations $t = 1,2,3,\ldots,k$ and let $y_{i1}, y_{i2}, ..., y_{ip}$ be the p-dimensional continuous random variable of interest with joint probability density function $f(y_i, \theta)$ on $\Re^P$ for $i = 1,2,3,...n$, where θ represents the vector of unknown parameters. The joint density of y_i can be defined as:

$$f(y_i, \theta) = \sum_{t=1}^{k} \pi_t f(y_i, \theta_t), \tag{1.1}$$

where $\pi_t = P(T = t)$, with $\sum_{t=1}^{k} \pi_t = 1$ and $\pi_t > 0 \forall t$, and where θ_t denote the vector of unknown parameters for cluster t. Each of the component density functions are assumed to come from a multivariate normal distribution parameterized by mean vector μ_t and variance covariance matrix Σ_t; that is, $\theta_t = (\mu_t, \Sigma_t)$. The unknown parameters are typically estimated using maximum likelihood, using an algorithm for finding the maximum of the likelihood such as the Expectation-Maximization algorithm (McLachlan and Peel, 2000). Various software packages implementing mixture of normals are currently available (e.g., Latent GOLD; Vermunt and Magidson 2005).

1.2.2 Classification Rules and Classification Errors

Once the cluster-specific parameters of the mixture distribution are estimated, the second step in the three-step approach involve allocating each subject to one of the k classes. We will denote the predicted class membership by W, with realization $s = 1,2,3,...,k$. The prediction for observation i is based on the cluster membership probabilities which can be obtained using Bayes' theorem:

$$P(T = t|y_i, \theta) = \frac{\pi_t f(y_i, \theta_t)}{f(y_i, \theta)}. \tag{1.2}$$

Let $w_{is} = P(W = s|y_i, \theta)$ be the likelihood of being assigned to class s given the assignment rule that is used. The most common rule is modal assignment, in which case w_{is} is a hard indicator; that is,

$$w_{is} = \begin{cases} 1 \text{ if } P(T = s|y_i, \theta) > P(T = t|y_i, \theta) \forall s \neq t \\ 0 \text{ otherwise} \end{cases} \tag{1.3}$$

An alternative rule is proportional assignment, in which case $w_{is} = P(T = s|y_i, \theta)$ (Vermunt, 2010).

Except for the situation in which $P(T = t|y_i, \theta)$ is either 0 or 1 for all i, there will be misclassifications. As discussed in more detail below, the total amount of classification errors can be quantified as the probability that a respondent belonging to cluster t is assigned to cluster s, which can be expressed as follows:

$$P(W=s|T=t)=\frac{P(W=s,T=t)}{P(T=t)} \tag{1.4}$$

The numerator of equation (1.4) is the joint marginal probability of W and T, which can be expressed in terms of the mixture model density. This yields:

$$\begin{aligned} P(W=s|T=t) &= \int \frac{P(W=s,T=t|y,\theta)f(y,\theta)dy}{P(T=t)} \\ &= \int \frac{P(T=t|y,\theta)P(W=s|y,\theta)f(y,\theta)dy}{P(T=t)}. \end{aligned} \tag{1.5}$$

The last step follows from the fact that W is independent of T conditional on y.

A complication factor in the computation of $P(W=s|T=t)$ is that the expression in equation (1.5) contain an intractable higher-dimensional integral. We propose solving this integral using Monte Carlo integration, which implies sampling say m units from $f(y,\theta)$ and computing the average of this sample. It should, however, be noted that if the mixture model holds, the sample used to solve integral can also be the n data points in the sample used to estimate the mixture model. This implies that $P(W=s|T=t)$ is approximated as follows:

$$P(W=s|T=t)\approx\frac{1}{n}\sum_{i=1}^{n}\frac{P(T=t|y_i,\theta)P(W=s|y_i,\theta)}{P(T=t)}. \tag{1.6}$$

1.3 Relationship Between Class Membership and Covariates

1.3.1 One-step Full Information ML Approach

Let z_i denote the vector with covariate values for subject i. In the one-step approach, inclusion of covariates involves expanding the standard mixture model defined in equation (1.1) as follows (Bandeen-Roche et al., 1997; Dayton and Macready, 1988):

$$f(y_i,\theta|z_i)=\sum_{t=1}^{k}P(T=t|z_i)f(y_i,\theta_t). \tag{1.7}$$

As can be seen, the prior class membership probabilities are now a function of covariates. These probabilities are typically modelled using a logistic regression equation; that is,

$$P(T=t|z_i)=\frac{\exp(\gamma_{0t}+\sum_{q=1}^{Q}\gamma_{qt}z_{iq})}{\sum_{m=1}^{k}\exp(\gamma_{0m}+\sum_{q=1}^{Q}\gamma_{qm}z_{iq})}.$$

The parameters of the mixture distribution and the covariate effects on the latent cluster membership can be estimated simultaneously using maximum likelihood estimation.

1.3.2 Standard Three-step Approach

An alternative is to use a three-step procedure. After estimating a standard mixture model and assigning individuals to classes, the relationship between the predicted class (W) and external variables is investigated using a standard multinomial logistic regression model:

$$P(W=s|z_i) = \frac{\exp(\gamma_{0s} + \sum_{q=1}^{Q} \gamma_{qs} z_{iq})}{\sum_{m=1}^{k} \exp(\gamma_{0m} + \sum_{q=1}^{Q} \gamma_{qm} z_{iq})}. \tag{1.8}$$

The γ parameters are estimated by maximizing the log-likelihood function:

$$\log L_{step3} = \sum_{i=1}^{n} \sum_{s=1}^{k} w_{is} \log P(W=s|z_i), \tag{1.9}$$

where in the case of modal assignment w_{is} is the hard indicator defined in equation (1.3).

1.3.3 Two Adjusted Three-step Approaches

The standard three-step approach defines a model for the relationship between external variables and the predicted cluster membership W instead of the true cluster membership T, which results in downward biased estimates for the covariate effects. However, Bolck et al. (2004) and Vermunt (2010) showed how to adjust for this bias by making use of the known relationship between $P(W=s|z_i)$ and $P(T=t|z_i)$.

More precisely, the adjustment methods described below are based on the following simple relationship:

$$P(W=s|z_i) = \sum_{t=1}^{k} P(T=t|z_i) P(W=s|T=t), \tag{1.10}$$

where $P(W=s|T=t)$ was defined in equations (1.4)-(1.6). It can be seen that $P(W=s|z_i)$ is a weighted sum of $P(T=t|z_i)$ where the $P(W=s|T=t)$ serve as weights.

The logic of the correction method proposed by Bolck, Croon, and Hagenaars (2004) – which we refer to as *the BCH approach* – is that if (1.10) holds, $P(T=t|z_i)$ can also be expressed as a weighted average of $P(W=s|z_i)$; that is,

$$P(T=t|z_i) = \sum_{s=1}^{k} P(W=s|z_i) d_{st}, \tag{1.11}$$

where d_{st} is an element of the inverse of the k-by-k matrix with elements $P(W = s|T = t)$. Bolck et al.(2004) proposed re-weighting the data on W (the class assignment weights w_{is}) by d_{st} to obtain approximate data on T. As shown by Vermunt (2010), the BCH approach can be implemented by creating an expanded data matrix containing k records per individual. The weight associated with the tth record equals $w^*_{it} = \sum_{s=1}^{k} w_{is} d_{st}$. A logistic regression model for T can now be estimated by maximizing the following weighted log-likelihood function:

$$\log L_{BCH} = \sum_{i=1}^{n} \sum_{t=1}^{k} w^*_{it} \log P(T = t|z_i). \tag{1.12}$$

Vermunt (2010) proposed using a sandwich variance estimator to take into account the weighting and the multiple observations per individual.

Vermunt (2010) proposed another simpler adjusted three-step method. It is based on the observation that equation (1.10) is in fact the equation of a latent class model with a single response variable W and with covariates. Since $P(W = s|T = t)$ is estimated step two, it can be treated as known in step three. Because this three-step approach involves maximizing a standard log-likelihood function in step three, we refer to it as *the ML approach*. More specifically, the parameters for the effects of the covariates on cluster membership can be estimated by maximizing the following log-likelihood function:

$$\log L_{ML} = \sum_{i=1}^{n} \sum_{s=1}^{k} w_{is} \log \sum_{t=1}^{k} P(T = t|z_i) P(W = s|T = t) \tag{1.13}$$

1.4 Simulation Study

1.4.1 Simulation Design

A simulation study was conducted to evaluate the performance of the various approaches for dealing with covariates in mixture models for continuous responses; i.e., the one-step ML, standard three-step, three-step BCH, and three-step ML method. Data sets were generated from a three-class mixture model for six continuous response variables and three covariates. The six responses were assumed to come from univariate normal distributions within classes. The residual variance was assumed to be equal across clusters and was used to manipulate the level of separation between clusters. The two independent factors that were manipulated were the separation between components, quantified using an entropy based R^2 measure, with low (R^2=.43), middle (R^2=.66) and high (R^2=.86) separation, and sample size (n=500; n=1000; n=10000). We looked at the averages of the covariate effects across replications, their standard deviations across replications, the averages of the standard errors of the estimates, and the mean square errors of the estimates. The Latent

Table 1.1 Results averaged over the nine combinations of sample size and separation between components for a γ parameter with a true value of 2

Model	Estimate	SE	SD	MSE
Standard	1.013	0.079	0.096	0.852
ML correction	1.962	0.210	0.223	0.231
One step	2.043	0.187	0.194	0.200
BCH	1.982	0.551	0.390	0.393

GOLD program (Vermunt and Magidson, 2005) was used in all stages of the simulation study such as generating data, estimating parameters in the various modeling approaches, getting classifications, and preparing expanded data sets. For this purpose, the program was called in a loop from a batch file.

1.4.2 Results

Table 1.1 presents the results averaged over all nine conditions for one of the covariate effect having a true value of 2. It can be seen that the standard three-step approach severely underestimates the parameter of interest, whereas both the ML and BCH correction method reduce the bias considerably. The correction methods drop the percentage of bias from 50% in the standard three-step approach to less than 2 %, which is similar to the bias of the one-step approach. The mean square error (MSE) of the estimates indicates that the ML correction method is almost as accurate as the one step method, whereas the BCH method is much less stable.

The performances of the various methods across the different simulation conditions were also investigated. The results reveal that except for the small sample size (n=500) and low separation (entropy=.43) combination, the BCH correction is found to have less bias than the ML correction and the one-step method. Consistent with the results of Table 1, the ML correction method is almost as efficient as the one-step method especially for the better separation and larger sample size conditions. In sum, the BCH method substantially reduces bias but is less efficient than the one-step and the ML correction method, while the ML correction provides estimates of a quality similar to the one-step approach for the more favorable conditions (larger sample sizes and higher separation levels).

1.5 Conclusions

This paper showed how to generalize the correction method for three-step latent class analysis with categorical response variables proposed by Bolck et al. (2004) and Vermunt (2010) to be applicable also in mixture models with continuous response variables. In agreement with theory on Monte Carlo integration, it was proposed to approximate the integral over the population density for the response vari-

ables by a summation over the observations in the data set at hand. What is clearly understood from the simulation results is that both the BCH and ML correction method performs quite well, except when the separation between components is extremely low and the sample size is small. This is in agreement with simulation results by Vermunt (2010) for mixture models with categorical responses. The practical advise to applied research is that one should not to use an uncorrected three-step method, but instead use one of the adjusted method. Only with extremely low separation levels combined with small samples, the one-step approach is clearly the best choice.

One issue requires further research, that is, finding an explanation for the instability of the BCH method and its overestimation of the SEs when used under the least favorable conditions. Further extensions of the correction methods would include situations where other categorical or continuous latent variables are used to explain the class membership, or more in general to any situation in which results from a mixture model clustering are used in subsequent analyses. These kinds of extension seems to be more straightforward with the ML method than with the BCH method.

References

1. Bandeen-Roche, K., Miglioretti, D.L., Zeger, S.L., and Rathouz, P.J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 92, 1375-1386.
2. Bolck, A., Croon, M.A., and Hagenaars, J.A.(2004). Estimating latent structural models with categorical variables: one –step versus three-step estimators. *Political Analysis*, 12, 3-27.
3. Dayton, C.M., and Macready, G.B. (1988). Concomitant-variable latent class models. *Journal of the American Statistical Association*, 83, 173-178.
4. McLachlan, G., and Peel, D. (2000). *Finite mixture models*. New York: John Wiley.
5. Vermunt, J.K.(2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18, 450-469.
6. Vermunt, J.K., and Magidson, J. (2005). *Latent GOLD 4.0 User's Guide*, Belmont, MA: Statistical Innovations Inc.