A reassessment of socio-economic gradients in child cognitive development using Growth Mixture Models

Supplementary Materials

Katy Sindall, *Department of Sociology, Social Policy, and Criminology, University of Southampton*

Patrick Sturgis, *Department of Social Statistics and Demography, University of Southampton*

Fiona Steele, *Department of Statistics, London School of Economics*

George Leckie, *Centre for Multilevel Modelling, University of Bristol*

Rob French, *School of Medicine, Cardiff University &* *Centre for Multilevel Modelling, University of Bristol*

DRAFT, PLEASE DO NOT CITE WITHOUT PRIOR PERMISSION

NOVEMBER 2018

Illustration of growth mixture models simulated data

**Motivation**

Using simulated data, we illustrate how GMMs recover features of the measurement model from which the data were generated. Perhaps the most compelling line of evidence in Jerrim and Vignoles’ (2013) critique of Feinstein (2003) is that they find the crossing pattern using the pre-assigned groups method when using simulated data, even though no such pattern existed in the simulated population from which the sample data were drawn. A key first step in assessing the utility of GMMs in this context then is to evaluate whether they can successfully recover the data generating mechanism, or whether the cross-over effect is also incorrectly produced.

**Data generation**

Growth mixture models (GMM) were applied to simulated data to assess how well they recover features of the measurement model from which the data were generated. The data were simulated using the same assumptions as Jerrim and Vignoles (2013). Denote by the ‘true ability’ at occasion for child and suppose that children are in one of two equally-sized SES groups indicated by such that for . Following Jerrim and Vignoles we assume the variance in child ability to be equal to 1 in each group, i.e. , and that a child’s ability is constant over time, i.e. for all . Suppose that is measured with error and that we observe which is related to by the linear measurement model

where the measurement error . Based on these assumptions, the data generating model (DGM) for can be written

 (S1)

where , and . Thus the measurement model takes the form of a random intercept multilevel model with a single predictor and no time trend.

We consider how the performance of the GMM depends on (i) the mean difference between the SES groups (), and (ii) the reliability of where . Following Jerrim and Vignoles, we generate data for each combination of = 1.4 and 3, and = 0.4 and 0.75, leading to four simulation conditions. Each simulated dataset contains 20,000 observations representing children, with each measured at all three occasions.

We fit two-class GMMs to the simulated data to investigate how well the latent classes map on to the observed SES categories. The fitted GMM has a more complex form than the DGM of equation (S1). In particular, we specified a model with a class-specific quadratic time trend (equivalent to including a dummy variable for each occasion for 3 time points), and allowed for class-specific between-individual (intercept) variances. Class and occasion-specific residual variances were also considered, but these models failed to converge. We therefore present results from models with the occasion-specific residual variances constrained to be equal across classes, i.e. for . Convergence problems were experienced with several of the models which is to be expected because the simplicity of the DGM means that more general specifications are not supported by the simulated data. Nevertheless it is important to release parameter constraints to assess the extent to which the parameter estimates are consistent with the DGM.

The fitted model is a simplification of the model fitted to the MCS data (equation (1) in the paper), excluding the random slope on and constraining the occasion specific residual variances to be class invariant:

where and . If the GMM perfectly reproduces the DGM, and the latent classes map on to the categories of , we would have , for , and . Furthermore, the average probability of class membership would be 0.5, .

The GMMs were fitted to standardised scores. While it is common to apply the pre-assigned groups method to percentiles, this is inappropriate for the GMM because percentiles follow a uniform distribution while a GMM assumes that residuals and random effects are normally distributed. Under the model of equation (S1) and using the result that for a binary variable with equal sized categories, the marginal variance of the observed responses is given by

 (S2)

Thus we compare estimates of with .

After estimating the GMM, we assessed the correspondence between the two latent classes and categories of by fitting a multinomial logistic regression of modal class membership , as in the three-step approach (e.g. Vermunt 2010; Asparouhov and Muthén, 2014), with as the only predictor. We anticipate that the association between and , as measured by the odds ratio , will be stronger when there is greater separation of the classes. Therefore we expect to be highest for the conditions where and .

**Results**

Table S1 shows the results from fitting two-class GMMs for three of the simulation conditions. The GMM did not converge for the fourth condition, and , which corresponds to the situation where the groups defined by are least well distinguished (i.e., small mean difference and low reliability). Figure S1 shows that in each case the predicted trajectories are constant, as in the DGM; thus and for the two classes are estimated close to zero in the GMM (results not shown). We also find that for each condition the estimated between-class difference in the intercepts is close to , the standardised mean difference assumed in the DGM (Table S1). Estimates of the intercept variances are also similar for the two classes, estimates of the residual variances , and are similar across the three occasions (Table S2), and estimates of the total variance within each class are close to the variance conditional on under the DGM (Table S3), . In terms of the correspondence between the latent classes and the observed SES indicator , the classes are of roughly equal size (Figure S1) and the substantial odds ratios show class membership is strongly associated with (Table S1). As expected, the classes are more closely aligned with SES when the groups are better distinguished.

TABLE S1 HERE

FIGURE S1 HERE

TABLE S2 HERE

TABLE S3 HERE

Although the results suggest that the GMM is able to recover the key features of the DGM, we note that entropy is low (Table S1) when the mean difference between SES categories is small or reliability is low. This may seem surprising given that the number of classes matches the number of groups in the DGM but entropy is a measure of the separation between classes and precision of the classification rather than model fit. It is possible for a correctly specified model to have a low entropy, as for two of the scenarios considered here (Muthén, 2004; Petras and Masyn, 2010). In such cases, the latent classes may not be useful and it is especially important to allow for uncertainty in class allocation when modelling the effects of covariates on class membership. Use of entropy-based fit indices (e.g. Celeux and Soromenho, 1996), might lead researchers to add spurious classes to the model in order to achieve more homogenous classes and therefore higher entropy. However, for all three simulation conditions, the addition of a third class led to non-convergence because, as expected given the DGM, the average probability of class membership for the additional class was very low (<0.5%).

PREDICTED PROBABILITY PLOTS OF CLASS MEMBERSHIP

FIGURES S2-8 HERE

**References**

Asparouhov, T. and Muthén, B. O. (2014). Auxiliary variables in mixture modeling: 3-step approaches using Mplus. *Mplus Web Notes No. 15*. Retrievable from www. statmodel. com.

Celeux, G. and Soromenho, G. (1996) An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification,* 13, 195-212.

Jerrim, J. and Vignoles, A. (2013) ‘Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes’, *Journal of the Royal Statistical Society Series A*, 176, 3.

Muthén, B. (2004). Latent variable analysis: growth mixture modeling and related techniques for longitudinal data. In: D. Kaplan (Ed.), *Handbook of Quantitative Methodology for the Social Sciences*, Sage Publications, CA, 345–368.

Petras, H. and Masyn, K. (2010) General growth mixture analysis with antecedents and consequences of change In *Handbook of Quantitative Criminology*  eds A. Piquero and D. Weisburd), pp. 69-100, New York: Springer.

Vermunt, J. K. (2010). Latent class modeling with covariates: two improved three-step approaches. *Political Analysis*, 18, 450-469.

Table S1. Results from fitting two-class GMMs to simulated data. The models assume quadratic growth and random intercepts within classes.

|  |  |  |
| --- | --- | --- |
| Simulation conditions | GMM estimates | Effect of on class membership |
| Mean group difference  | Reliability  | Standardised group difference  |  | Entropy | Odds ratio | p-value |
| 1.4 | 0.75 | 0.932 | 0.988 | 0.275 | 3.66 | <0.001 |
| 3 | 0.75 | 1.494 | 1.499 | 0.695 | 44.61 | <0.001 |
| 3 | 0.40 | 1.059 | 1.075 | 0.456 | 9.67 | <0.001 |

Note: and are estimates of the intercepts for each class in the GMM.

Table S2. Estimates (and standard errors) of intercept and residual variances from fitting two-class GMMs to simulated data. The models assume quadratic growth and random intercepts within classes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| ,  |
| 1 | 0.418 | (0.058) | 0.348 | (0.006) | 0.343 | (0.006) | 0.360 | (0.005) |
| 2 | 0.396 | (0.059) |
| ,  |
| 1 | 0.237 | (0.010) | 0.197 | (0.003) | 0.194 | (0.003) | 0.202 | (0.003) |
| 2 | 0.240 | (0.010) |
| ,  |
| 1 | 0.128 | (0.019) | 0.594 | (0.009) | 0.585 | (0.009) | 0.612 | (0.009) |
| 2 | 0.107 | (0.019) |

Note: is the group mean difference and is the reliability assumed in the data generation model. Tables S1 and S2 present parameter estimates from the same models.

Table S3. Comparison of estimates of the within-class variance from the GMM with the conditional variance of the standardised response given under the data generating model.

|  |  |
| --- | --- |
| Simulation conditions | Total within-class variance  |
| Mean group difference  | Reliability  |  |  |  |
| 1.4 | 0.75 | 0.78 | 0.75 | 0.77 |
| 3 | 0.75 | 0.44 | 0.44 | 0.44 |
| 3 | 0.40 | 0.74 | 0.72 | 0.70 |

Notes: Under the data generating model, the conditional variance of the standardised response given is , from equation (S2). The calculation of the within-class variance uses the mean of the residual variances across occasions, .

Figure S1. Predicted trajectories and latent group membership probabilities from 2-class GMM models fitted to simulated data. Fitted models assume quadratic mean growth and random intercepts within latent classes.

1. (b) (c)

   

   

Figure S2. Line plot of the predicted probability of class membership against standardised income holding all other covariates at their mean values. The percentages reported in the legend state the overall percentage of individuals in each class.



Figure S3. Bar chart of the predicted probability of class membership against benefit payments holding all other covariates at their mean values. The percentages reported in the legend state the overall percentage of individuals in each class.



Figure S4. Bar chart of the predicted probability of class membership against NS-SEC holding all other covariates at their mean values. The percentages reported in the legend state the overall percentage of individuals in each class.



Figure S5. Bar chart of the predicted probability of class membership against marital status holding all other covariates at their mean values. The percentages reported in the legend state the overall percentage of individuals in each class.



Figure S6. Bar chart of the predicted probability of class membership against parent long term illness holding all other covariates at their mean values. The percentages reported in the legend state the overall percentage of individuals in each class.



Figure S7. Bar chart of the predicted probability of class membership against parent’s age at birth holding all other covariates at their mean values. The percentages reported in the legend state the overall percentage of individuals in each class.



Figure S8. Bar chart of the predicted probability of class membership against child’s gender holding all other covariates at their mean values. The percentages reported in the legend state the overall percentage of individuals in each class.

