A two-level multivariate response model for data with latent structures

Yingjuan Zhang[1], Jochen Einbeck[1,2] and Reza Drikvandi[1]
[1]Department of Mathematical Sciences, Durham University, Durham, UK
[2]Durham Research Methods Centre, Durham University, Durham, UK

Abstract: A novel approach is proposed for analysing multilevel multivariate response data. The approach is based on identifying a one-dimensional latent variable spanning the space of responses, which then induces correlation between upper-level units. The latent variable, which can be thought of as a random effect, is estimated along with the other model parameters using an EM algorithm, which can be seen in the tradition of the 'nonparametric maximum likelihood' estimator for two-level linear (univariate response) models. Simulations and real data examples from different fields are provided to illustrate the proposed methods in the context of regression and clustering applications.

1 Introduction

When data possess a repeated measures structure, such as pupils nested within schools, or longitudinal measurements of individuals over time, the use of random effects to account for the ensuing correlations is now a commonplace technique. Specifically, the idea is to equip each upper-level unit with a random intercept which is shared by all lower-level units pertaining to it, and which induces the required correlations.

While these methods are well-developed and well-understood, and well supported by statistical software, they are typically restricted to univariate response scenarios. When the space of responses is multivariate, it is however not very clear how to actually adopt the idea mentioned above: Should there be a single or multiple random effects per upper-level unit, or in other words, shall the random effect distribution also be multivariate, and in either case, what is the shape of this distribution and how to estimate its parameters?

Before we begin with presenting our answer to these questions, we will briefly outline three examples for such problems which will serve as case studies later in this exposition. First in research

Address for correspondence: Yingjuan Zhang, Department of Mathematical Sciences, Durham University, Stockton Road, Durham, UK.
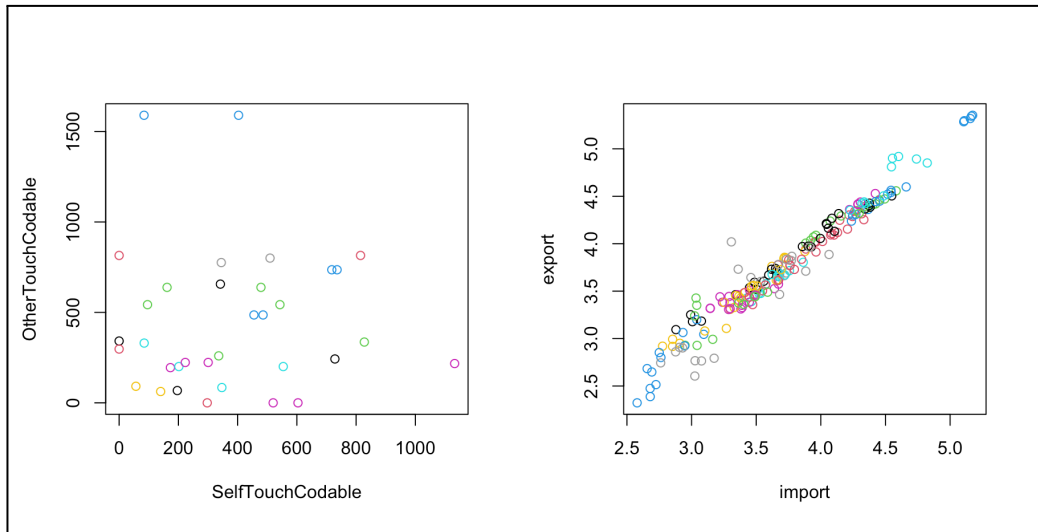E-mail: yingjuan.zhang@durham.ac.uk

Figure 1 Left: Fetal twins' touch movements data, coloured by mothers. Right: Import and export data, coloured by countries.

on the effects of maternal mental health on prenatal movements in fetuses (Reissland et al., 2021), two touch movement types of twin fetuses were recorded during 4D ultrasound scans: self touch (the fetus touching itself) and other touch (the fetus touching the other twin). Figure 1 (left) shows the scatter plot of the two response variables symbolized by values of the upper-level variable 'mother'. The objective is to investigate the effect of maternal mental health (depression, stress and anxiety) onto the movement profile of twins, taking the correlation of the measurements of fetuses belonging to the same mother into account.

Second we consider data from the OECD (Organisation for Economic Co-operation and Development, 2023b) concerning trade in goods and services, providing country-wise percentages of imports and exports in relation to the overall GDP in 44 countries, for the time period between 2018 and 2022, during which between 3 and 5 observations are available for each country. Figure 1 (right) visualizes the data where the observations from the same country have the same colour. This can be considered as a multivariate repeated measures scenario, with unbalanced measurement occasions, and without covariates. We are particularly interested in clustering the countries with respect to their overall export/import activity relative to GDP size, taking within-country correlations across the repeated measurements into account. In recent related work, albeit employing a different methodology involving hidden Markov models, Pennoni et al. (2024) propose an algorithm for selecting the most important variables to cluster and classify countries by socio-economic development.

Third we consider the Programme for the International Assessment of Adult Competencies (PIAAC) survey of adult skills, carried out in 2011 and 2012 by the OECD. The PIAAC survey was designed to assess the proficiency of adults in the key information-processing skills of literacy, numeracy and problem solving (in technology-rich environments). All three skill types are provided on a continuous scale ranging from 0 to 500. For our analysis, we extracted from the PIAAC explorer (https://piaacdataexplorer.oecd.org/ide/idepiaac/) data from 28 countries and two

Figure 2 Pairs plot of PIAAC data, coloured by countries.

sub-national regions on all three criteria with two covariates: gender and current work status (employee or self employed). Figure 2 shows the correlation between the three response variables, each plotted against the others and coloured by the upper-levels (countries). As in the previous example, we are interested in the clustering of countries in the presence of country-level correlations, with the focus shifting here towards the creation of a league table of countries. A secondary interest lies in the study of the effect of the covariates on the outcomes. A somewhat similar analysis (using Stata) was carried out by Grilli et al. (2016) using data from the TIMSS&PIRLS database. Their multivariate approach jointly considers educational achievement in reading, mathematics and science, where the coefficients for each response were estimated separately and combined using multiple imputation formulas. However, they did not consider the ranking problem, and their approach cannot be used for clustering purposes.

We provide a modelling approach which will allow us to tackle the problems above, taking into account both the multivariate and the multilevel character of the outcome data. The approach is based on Zhang and Einbeck (2024a)'s latent variable model for dimension reduction and simultaneous clustering of highly correlated data, requiring only a single, one-dimensional random effect term. This paper develops the upheaval of that approach to two-level scenarios which is required to be able to deal with repeated measures data. We will give equal importance to the applications of clustering (of upper-level units) and multivariate regression for two-level data.

Some related model classes have been developed in the wider context of item response theory, most notably latent class models (Goodman, 1974). These models are commonly used for the clustering of observed multivariate categorical data (such as questionnaire outcomes on Likert scales) into latent classes. An obvious difference to our methodology is that in latent class models

the response variables are categorical rather than continuous. A multilevel version of latent class models was developed by Vermunt (2003). A model selection procedure for deciding the number of latent classes at both levels is proposed by Lukočienė et al. (2010). Gnaldi et al. (2016) introduced a multilevel version latent class-item response theory model applied for educational data in which the collected response variables are dependent of each other. The latent class analysis also allows the inclusion of covariates; Di Mari et al. (2023) proposed a two-step estimator for the multilevel latent class model in which two categorical random effects are used to account for both the upper and lower-levels, allowing for clustering of the latent classes on both levels. It remains the case that due to the restriction on categorical outcomes, latent class models cannot be applied or compared with the situations dealt with in this work. However, it should not be left unstated that continuous-outcome versions of multilevel latent class models have also been developed, and are available in specialized commercial software such as Latent GOLD (Vermunt, 2008). Further related work includes Masci et al. (2022) who proposed a semiparametric mixed-effects model for multinomial data with hierarchical structure, in which a discrete random effect distribution is used to obtain the marginal density, and Bartolucci et al. (2011) who proposed a multilevel extension of latent Markov Rasch model and applied this on educational data with three-level structures. Verbeke et al. (2014) gave a general overview over longitudinal models for multivariate outcome data.

The structure of this paper is as follows. In Section 2 we introduce the proposed two-level model for multivariate response data. In Section 3, we present an EM algorithm for the proposed model, resembling the nonparametric maximum likelihood method. Section 4 shows simulation results that demonstrate the performance and accuracy of this algorithm for the estimation of model parameters. Section 5 provides real data examples that illustrate the main applications of our model, including the fitting of a multivariate response model resulting in reduced standard errors, the construction of league tables and the clustering of upper-level units based on the fitted model.

Some additional simulation results and complementary information have been relegated to the supplementary material. R Codes of the implemented methods, as well as of several of the presented examples, are available in R package mult.latent.reg, which is available on CRAN (Zhang and Einbeck, 2024b).

2 A two-level model for multivariate response data

We consider a scenario where multivariate data $x_{ij} \in \mathbb{R}^m$ have a two-level structure, with the upper-level indexed by $i = 1, 2, ..., r$ and the lower-level by $j = 1, 2, ..., n_i$. The proposed two-level model takes the form

$$x_{ij} = \alpha + \beta z_i + \Gamma v_{ij} + \varepsilon_{ij}, \tag{2.1}$$

where $\alpha, \beta \in \mathbb{R}^m$, $z_i \in \mathbb{R}$, $v_{ij} \in \mathbb{R}^p$ is the vector of covariates (which may include upper-level variates not depending on j), $\Gamma \in \mathbb{R}^{m \times p}$ is a matrix of the covariate coefficients, and $\varepsilon_{ij} \sim N(0, \Sigma(z_i))$ are independent Gaussian errors. Under such a model, equivalently represented as

$$x_{ij}|z_i, \alpha, \beta, \Gamma \sim N(\alpha + \beta z_i + \Gamma v_{ij}, \Sigma(z_i)), \tag{2.2}$$

the data grouping process is carried out on the upper-level, while the lower-level units within the same upper-level unit share a common random effect term z_i. Thus, the random effect induces a line

cutting across the multivariate space of responses, along which the latent values z_i are positioned. Again equivalently, and for later reference, we can write the conditional probability density function of the x_{ij} as

$$f(x_{ij}|z_i,\alpha,\beta,\Gamma) = \tag{2.3}$$
$$(2\pi)^{-m/2}|\Sigma(z_i)|^{-1/2}\exp\left\{-\frac{1}{2}(x_{ij}-\alpha-\beta z_i-\Gamma v_{ij})^T\Sigma^{-1}(z_i)(x_{ij}-\alpha-\beta z_i-\Gamma v_{ij})\right\}.$$

For the distribution of random effects z_i, denoted here by Z, several choices are possible, including a Gaussian distribution. In this work, we consider to use Aitkin's nonparametric maximum likelihood approach (Aitkin, 1999), in which their distribution is approximated by a discrete mixture. However, as will be detailed in the following section, this is not so much a distributional 'assumption', but rather a technical device to approximate the marginal likelihood, allowing for estimation of the model parameters. De facto this approach leads to the estimation of a constrained multivariate mixture model, with mixtures centres spanned along a straight line through the space of responses. When there is only one covariate $v_{ij} \in \mathbb{R}$, we write $\Gamma = \gamma \in \mathbb{R}^m$. Figure S1 in part A of the supplementary material gives a graphical illustration of how the model operates in this case.

3 Methods and estimation

3.1 Likelihood and estimators

Let $x_i = (x_{i1},\ldots,x_{in_i})^T \in \mathbb{R}^{n_i \times m}$ denote the collection of the m-variate lower-level observations relating to the ith upper-level unit. Since these lower-level units are conditionally independent given z_i, we have

$$f(x_i|z_i,\alpha,\beta,\Gamma) = \prod_{j=1}^{n_i} f(x_{ij}|z_i,\alpha,\beta,\Gamma).$$

According to model (2.1), the marginal distribution of x_i, which is required for the construction of the likelihood function, can be obtained by integrating over the distribution of z_i, as follows:

$$f(x_i|\alpha,\beta,\Gamma) = \int \left[\prod_{j=1}^{n_i} f(x_{ij}|z_i,\alpha,\beta,\Gamma)\right] g(z_i)dz_i, \tag{3.1}$$

where $g(z_i)$ is the density function for the unobserved random effects z_i. Under the nonparametric maximum likelihood approach (Aitkin, 1999), we replace the integral over z_i by a finite sum over K mass points $z_1,\ldots,z_k$ with associated masses $\pi_1,\ldots,\pi_k$, for $k = 1,\ldots,K$. Here we treat the mass points and masses as unknown parameters to be estimated. The value of K will be treated as known in the parameter estimation process and the best choice of K in a fitted model will be selected through the use of model selection criteria, specifically based on the AIC criterion.

The marginal distribution can then be approximated as

$$f(x_i|\alpha,\beta,\Gamma) \approx \sum_{k=1}^{K} \left[\prod_{j=1}^{n_i} f(x_{ij}|z_k,\alpha,\beta,\Gamma) \right] \pi_k, \tag{3.2}$$

in which, by virtue of (2.2),

$$x_{ij}|z_k,\alpha,\beta,\Gamma \sim N(\alpha + \beta z_k + \Gamma v_{ij}, \Sigma(z_k)), \tag{3.3}$$

with the component-specific densities $f(x_{ij}|z_k,\alpha,\beta,\Gamma)$ as in equation (2.3), but with z_i replaced by z_k.

Now, the $\alpha + \beta z_k$ can be interpreted as the locations, in m-dimensional space, of the mixture centres spanned along the one-dimensional latent space, with cluster-wise variances $\Sigma_k \equiv \Sigma(z_k)$ replacing the previous observation-specific variances $\Sigma(z_i)$. The number of parameters to be estimated is effectively reduced by constraining to K distinct variance matrices.

Building on equation (3.2), the approximated marginal log-likelihood can be obtained as

$$l(\alpha,\beta,\Gamma,z_1,\ldots,z_K|x_1,\ldots x_r) \approx \sum_{i=1}^{r} \log \left\{ \sum_{k=1}^{K} \left[\prod_{j=1}^{n_i} f(x_{ij}|z_k,\alpha,\beta,\Gamma) \right] \pi_k \right\}. \tag{3.4}$$

In preparation of the EM algorithm (e.g., Dempster et al., 1977) to be used for the parameter estimation, we define by G_{ik} an indicator variable taking the value 1 if the upper-level unit i belongs to component k, and 0 otherwise (which is, of course, unknown—this is the 'missing information' for the EM machinery). We also denote by $G_i = (G_{i1},\ldots,G_{iK})^T$ the set of indicators for that unit. This yields 'complete data' $\{x_i, G_i\}$, with probability

$$P(x_i, G_i) = \prod_{k=1}^{K} (f_{ik}\pi_k)^{G_{ik}},$$

where for simplicity of notation we here used $f_{ik} \equiv \prod_{j=1}^{n_i} f(x_{ij}|z_k,\alpha,\beta,\Gamma)$. The complete likelihood can now be written as follows:

$$L_c = \prod_{i=1}^{r} \prod_{k=1}^{K} (\pi_k f_{ik})^{G_{ik}}. \tag{3.5}$$

Hence, we obtain the complete log-likelihood,

$$l_c = \log L_c = \sum_{i=1}^{r} \sum_{k=1}^{K} G_{ik} \log(\pi_k f_{ik}) \tag{3.6}$$

which may take values in $(-\infty, \infty)$. The expectation $w_{ik} = E[G_{ik}|x_i] = P(G_{ik} = 1|x_i) = \pi_k f_{ik} / \sum_\ell \pi_\ell f_{i\ell}$ is just the 'posterior' probability of each upper-level unit i belonging to component

k. Therefore, the expected complete log-likelihood is written as

$$l_c^* = \sum_{i=1}^{r}\sum_{k=1}^{K}\mathbb{E}\left[G_{ik}|x_i\right]\log\left(\pi_k f_{ik}\right)$$

$$= \sum_{i=1}^{r}\sum_{k=1}^{K}w_{ik}\log\pi_k + \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K}w_{ik}\log f(x_{ij}|z_k,\alpha,\beta,\Gamma).$$

(3.7)

Plugging the expression for $f(x_{ij}|z_k,\alpha,\beta,\Gamma)$ into equation (3.7), we obtain the expected complete log-likelihood as follows:

$$l_c^* = \sum_{i=1}^{r}\sum_{k=1}^{K}w_{ik}\log(\pi_k) - \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K}w_{ik}\log(|\Sigma_k|) - \frac{m}{2}\log(2\pi)\sum_{i=1}^{r}n_i$$

$$- \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K}w_{ik}(x_{ij}-\alpha-\beta z_k - \Gamma v_{ij})^T\Sigma_k^{-1}(x_{ij}-\alpha-\beta z_k - \Gamma v_{ij}).$$

(3.8)

By taking partial derivatives of l_c^* with respect to each parameter and letting the score equations to be 0 and solving them, we find that

$$\hat{\alpha} = \left(\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K}w_{ik}\hat{\Sigma}_k^{-1}\right)^{-1}\left(\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K}w_{ik}\hat{\Sigma}_k^{-1}(x_{ij}-\hat{\beta}\hat{z}_k - \hat{\Gamma}v_{ij})\right),$$

(3.9)

$$\hat{\beta} = \left(\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K}w_{ik}\hat{\Sigma}_k^{-1}\hat{z}_k^2\right)^{-1}\left(\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K}w_{ik}\hat{\Sigma}_k^{-1}(x_{ij}-\hat{\alpha}-\hat{\Gamma}v_{ij})\hat{z}_k\right),$$

(3.10)

$$\hat{z}_k = \frac{\sum_{i=1}^{r}w_{ik}\sum_{j=1}^{n_i}\hat{\beta}^T\hat{\Sigma}_k^{-1}(x_{ij}-\hat{\alpha}-\hat{\Gamma}v_{ij})}{\hat{\beta}^T\hat{\Sigma}_k^{-1}\hat{\beta}\sum_{i=1}^{r}n_iw_{ik}}, \, k = 1,\ldots,K.$$

(3.11)

The solution for $\hat{\Gamma}$ can only be given implicitly in the form of estimating equation

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K}w_{ik}\hat{\Sigma}_k^{-1}(x_{ij}-\hat{\alpha}-\hat{\beta}z_k)v_{ij}^T = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K}w_{ik}\hat{\Sigma}_k^{-1}\hat{\Gamma}v_{ij}v_{ij}^T.$$

(3.12)

We furthermore find the general solution for Σ_k as

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i}w_{ik}(x_{ij}-\hat{\alpha}-\hat{\beta}\hat{z}_k - \hat{\Gamma}v_{ij})(x_{ij}-\hat{\alpha}-\hat{\beta}\hat{z}_k - \hat{\Gamma}v_{ij})^T}{\sum_{i=1}^{r}n_iw_{ik}},$$

(3.13)

for $k = 1,\ldots,K$. Finally, since for the mixture probabilities $\sum_{k=1}^{K}\pi_k = 1$, we apply a Lagrange multiplier by letting $\partial\left(l_c^* - \lambda(\sum_{k=1}^{K}\pi_k - 1)\right)/\partial\pi_k = 0$, with Lagrangian parameter $\lambda \in \mathbb{R}$. Hence,

we find

$$\hat{\pi}_k = \frac{\sum_{i=1}^{r} w_{ik}}{r}. \tag{3.14}$$

We note that this set of equations (3.9) to (3.14) is rather impractical to use directly, because the equations depend on each other in a complex manner, they involve multiple inversions of the estimated matrices $\hat{\Sigma}_k$, and the solution for Γ does not have an explicit form. However, it is also not necessary to apply these equations in full generality. An immediate simplification is suggested by considering the matrices Σ_k. While these variance matrices, under a full unconstrained parameterization, could deal with clusters that differ by shape and size, when fitting a multi-level model, the focus is unlikely to be on estimating the shape of the clusters. Hence, we will restrict to diagonal variance matrices

$$\Sigma_k = \text{diag}(\sigma_{lk}^2)_{\{1 \le l \le m\}}, k = 1, ..., K.$$

To avoid potential identifiability issues, certain restrictions are imposed on the model. First, we enforce $\beta_1 \ge 0$ to identify the direction of the latent variable. Then we standardize z_k by $\sum_{k=1}^{K} \pi_k z_k = 0$, and $\sum_{k=1}^{K} \pi_k z_k^2 - (\pi_k z_k)^2 = 1$, where $\text{Var}[z_k] = \sum_{k=1}^{K} \pi_k z_k^2 - (\pi_k z_k)^2$ (Marques da Silva Júnior et al., 2018).

The resulting EM algorithm, which makes some further simplifications which are however of computational rather than model-related character, is presented in the next subsection.

3.2 EM algorithm

We have the following expectation (E) and maximization (M) steps resulting from the previous considerations.

E-step: The E-step is obtained from the straightforward application of Bayes' theorem as illustrated in the previous subsection,

$$w_{ik} = \frac{\pi_k f_{ik}}{\sum_l \pi_l f_{il}}. \tag{3.15}$$

M-step: In order to implement the M-step computationally, we adopt the strategy employed in Zhang and Einbeck (2024a). For this, we detach the updates of $\hat{\alpha}$, $\hat{\beta}$, $\hat{z}_k$ and $\hat{\Gamma}$ from those of $\hat{\Sigma}_k$, by invoking, only for the use within expressions (3.9) to (3.12), a further simplification where the variance matrices are assumed to be constant and diagonal, i.e. $\sigma_{lk}^2 \equiv \sigma^2$ for all l and k. This leads to simpler equations for (3.9) to (3.12) as follows:

$$\hat{z}_k = \frac{\sum_{i=1}^{r} w_{ik} \sum_{j=1}^{n_i} \hat{\beta}^T (x_{ij} - \hat{\alpha} - \hat{\Gamma} v_{ij})}{\hat{\beta}^T \hat{\beta} \sum_{i=1}^{r} n_i w_{ik}}, \tag{3.16}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{r} \sum_{j=1}^{n_i} \sum_{k=1}^{K} w_{ik} \hat{z}_k x_{ij} - \frac{1}{n} (\sum_{i=1}^{r} \sum_{j=1}^{n_i} x_{ij})(\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik} \hat{z}_k)}{\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik} \hat{z}_k^2 - \frac{1}{n} (\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik} \hat{z}_k)^2}$$
$$- \frac{\hat{\Gamma} \sum_{i=1}^{r} \sum_{j=1}^{n_i} \sum_{k=1}^{K} w_{ik} \hat{z}_k v_{ij} - \frac{1}{n} (\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik} \hat{z}_k)(\hat{\Gamma} \sum_{i=1}^{r} \sum_{j=1}^{n_i} v_{ij})}{\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik} \hat{z}_k^2 - \frac{1}{n} (\sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik} \hat{z}_k)^2},$$

$$\hat{\alpha} = \frac{1}{n} \left(\sum_{i=1}^{r} \sum_{j=1}^{n_i} x_{ij} - \hat{\beta} \sum_{i=1}^{r} n_i \sum_{k=1}^{K} w_{ik} \hat{z}_k - \hat{\Gamma} \sum_{i=1}^{r} \sum_{j=1}^{n_i} v_{ij} \right),$$

with the estimator for Γ now being available in explicit form,

$$\hat{\Gamma} = \left(\sum_{i=1}^{r} \sum_{j=1}^{n_i} v_{ij} v_{ij}^T \right)^{-1} \left(\sum_{i=1}^{r} \sum_{j=1}^{n_i} \sum_{k=1}^{K} w_{ik} (x_{ij} - \hat{\alpha} - \hat{\beta} \hat{z}_k) v_{ij}^T \right).$$

These four equations are then iterated for a small number of times between each other, where the $\hat{z}_k, k = 1, \ldots, K$, are immediately re-standardized to mean 0 and variance 1 after the execution of (3.16) . This routine is then followed by the estimation of the π_k via (3.14), and the update of the variance matrices via $\hat{\Sigma}_k = \text{diag}(\hat{\sigma}_{lk}^2)_{\{1 \leq l \leq m\}}, k = 1, ..., K$. Write $\phi_{ij} = \hat{\Gamma} v_{ij} \in \mathbb{R}^m$ and let $\phi_{ij\ell}$ be its ℓth component, $\ell = 1, \ldots, m$. Then

$$\hat{\sigma}_{lk}^2 = \frac{\sum_{i=1}^{r} \sum_{j=1}^{n_i} w_{ik} (x_{ijl} - \hat{\alpha}_l - \hat{\beta}_l \hat{z}_k - \phi_{ijl})^2}{\sum_{i=1}^{r} n_i w_{ik}}.$$

This completes the M-step, and the procedure continues with the E-step (3.15).

Several options for selecting starting values have been implemented in the R package mult.latent.reg, which we described in Zhang and Einbeck (2024c). The required number of iterations is usually small, so that automated assessment of convergence is not necessary. The implementation in mult.latent.reg uses by default 20 iterations, which is generally sufficient. It is worth noting that due to the sequential nature of the updates within the M-step, this algorithm can be considered an ECM algorithm, for which convergence is, however, still guaranteed (Meng and Rubin, 1993). Choosing the number of mixture components K is a model selection process. Here, we use the AIC $= -2l + 2q$, where l is the log-likelihood, defined using equation (3.4), and $q = 2(K - 1) + m(2 + K + p)$ is the total number of parameters.

4 Simulation studies

4.1 Evaluate the accuracy of parameter estimation

We first conduct a simulation study to examine the accuracy of our parameter estimation using the simplified update expressions for the EM algorithm as described in Section 3.2. Another objective of this simulation is to investigate whether an increase in the number of upper- or lower-level units will effectively reduce the variance of the parameter estimates. We simulate data from bivariate two-level scenarios with a single covariate, where the number of mixture components is $K = 2$. We first consider a scenario with $r = 50$ upper-level units and $n_i = 5$ lower-level units, for $i = 1, 2, \ldots, r$. This will be the baseline experiment. Then we keep $r = 50$ unchanged and increase the number of lower-level units to be $n_i = 10$, for $i = 1, 2, \ldots, r$. We consider another sample size with lower-level units $n_i = 5$ for $i = 1, 2, \ldots, r$ unchanged but increase the upper-level units to be $r = 100$. We also further increase the upper-level units to be $r = 200$ and keep the lower-level units $n_i = 5$ for $i = 1, 2, \ldots, r$. We generate 200 replicated data sets (each with two mixture components, $\pi_1 = 0.4$,

Table 1 Estimates of key parameters γ, z_k and α with different numbers of upper-level and lower-level units.

	True	Average estimates			
		$r = 50, n_i = 5$	$r = 50, n_i = 10$	$r = 100, n_i = 5$	$r = 200, n_i = 5$
γ_1	1.000	1.033	0.981	0.990	0.997
γ_2	3.000	3.031	3.034	2.993	3.004
z_1	-0.816	-0.804	-0.815	-0.818	-0.814
z_2	1.225	1.279	1.256	1.236	1.235
α_1	2.000	1.986	2.041	2.022	1.990
α_2	10.000	9.995	10.021	10.007	10.001

Table 2 RMSE for key parameters γ, z_k and α with different numbers of upper-level and lower-level units.

	RMSE			
	$r = 50, n_i = 5$	$r = 50, n_i = 10$	$r = 100, n_i = 5$	$r = 200, n_i = 5$
γ_1	0.278	0.157	0.166	0.111
γ_2	0.441	0.284	0.263	0.201
z_1	0.130	0.125	0.082	0.057
z_2	0.233	0.200	0.129	0.087
α_1	0.455	0.429	0.310	0.213
α_2	0.179	0.157	0.116	0.077

$\pi_2 = 0.6$ and true values of z_k's as shown in the first column of Table 1) from the model (3.3). In all four scenarios a lower-level covariate is generated from a normal distribution with mean 0.3 and standard deviation 0.2, and with true $\gamma = (1, 3)^T$.

For the estimation from the simulated data, we also use $K = 2$. The effect of misspecifying K is considered in the supplementary materials. The simulation results, which are presented in Tables 1 and 2 and Figure 3, indicate that the true parameters are well estimated, and when we increase the number of upper-level units, the parameters' RMSE decreases stronger than when increasing the number of lower-level units. Note that, for a univariate parameter θ, the root mean squared error is defined as RMSE $= \sqrt{\frac{\sum_{i=1}^{s}(\theta_0 - \hat{\theta}_i)^2}{s}}$, where θ_0 is the true value, $\hat{\theta}_i$ is the ith estimated value, and s is the number of simulation runs (so, here $s = 200$).

We also compare the γ estimates from our model to those obtained by fitting individual two-level models. Each of these models uses one of the simulated two-dimensional variables as response variable and treats the covariate as predictor. We used the lmer() function in R package lme4 and the allvc() function from the npmlreg package for this comparison. The results, displayed in Tables 3 and 4, show that our method produces sensible results when compared to those obtained with allvc() and even superior estimates when compared to those obtained with lmer().

4.2 Further simulation studies

Further analyses concerning the impact of misspecification of the number of mixture components and the random effect distribution are relegated to the supplementary material B and C. In brief,

Figure 3 Estimates of key parameter γ with different number of upper-level and lower-level units.

Table 3 Averaged estimates of γ obtained by fitting individual models to each response variable.

			Average estimates		
	True	$r = 50, n_i = 5$	$r = 50, n_i = 10$	$r = 100, n_i = 5$	$r = 200, n_i = 5$
			lmer()		
γ_1	1.000	0.999	0.987	0.989	0.996
γ_2	3.000	2.972	3.037	3.002	2.999
			allvc()		
γ_1	1.000	0.993	0.992	0.989	0.995
γ_2	3.000	2.998	3.037	3.005	2.995

Table 4 RMSE for γ obtained by fitting individual models to each response variable.

			RMSE	
	$r = 50, n_i = 5$	$r = 50, n_i = 10$	$r = 100, n_i = 5$	$r = 200, n_i = 5$
			lmer()	
γ_1	0.286	0.182	0.175	0.123
γ_2	0.470	0.325	0.278	0.209
			allvc()	
γ_1	0.259	0.167	0.166	0.115
γ_2	0.396	0.284	0.263	0.191

these simulation results confirm that the regression parameter γ is unaffected by the number of components, but that it is slightly affected by the random effect distribution, particularly if that distribution is continuous and skewed.

Table 5 The values of AIC$=-2\ell + 2q$ for the twins data fitted with different number of mixture components. The best solution is highlighted in bold.

K	1	2	3	4	5	6
ℓ	-203.772	-197.313	-194.307	-193.179	-193.202	-192.425
q	13	17	21	25	29	33
AIC	433.545	**428.627**	430.615	436.359	444.403	450.851

5 Analyses of case studies

In this section we analyse the real data sets from our case studies briefly introduced in Section 1. We focus on regression in the first case study and on clustering in the second case study, while in the third case study both regression and clustering are of interest.

5.1 Fetal twins' touch movements

The data set considered here was originally collected for research on the effects of maternal mental health on prenatal movements in twins and singletons (see Reissland et al., 2021). Since we are interested in a joint modelling of the two touch dimensions 'self touch' and 'other touch', we work here with slightly reduced data where the singletons are omitted (because singletons can't touch the 'other' twin). In the remaining twins' data, from 14 mothers who were pregnant with twins, 11 mothers were available for one scan and 3 were available for two scans, that is, in total there are 34 observations. Besides the two touch movement types, at the ultrasound scan appointment, the mothers' mental health status was collected on three variables: depression, perceived stress scale and anxiety. The data set as used in this case study is available as twins_data from R package mult.latent.reg.

For our analysis of the twins data set, we consider the two types of touches, self touch and other touch, as a bivariate response, and include the three mental health variables as covariates into model (2.1). Under this model, the observations within upper-levels (we consider each mother as an upper-level unit) share a common, mother-specific, random effect z_i, which accounts for correlated touch behaviour of fetuses from the same mother. Notably there is only one such random effect variable, which applies to both response variables.

An examination of the AIC values across different values of K (Table 5) yields that the minimum AIC is attained for $K = 2$, with AIC value 428.627, and hence we use this choice of K for our analysis. The traditional method of dealing with such data would be fitting separate two-level models, each using one of the touch movements as the response variable and the three mental health measurements as covariates. Table 6 shows the estimates of the coefficients and their standard errors obtained through using the lmer() function in R package lme4 (Bates et al., 2015), and Table 7 shows the estimates from our model and the bootstrapped standard errors. Note that the bootstrap applied here is a straightforward extension of the bootstrap technique developed by Zhang and Einbeck (2024a), adjusted to the context of the two-level models, ensuring that all units on the upper-level get associated with the same random effect. Our approach gives reduced standard errors compared to the linear mixed model, with similar parameter estimates (in the sense of, being comfortably within one standard error of the respective other model).

Table 6 For the twins data, estimations of γ obtained using individual two-level models (lmer()) for self touch and other touch as response and depression, perceived stress scale (PSS) and anxiety as predictors, with standard errors given in brackets.

	Indiv. two-level models		
	Depression	Stress	Anxiety
Self touch	−27.34 (39.18)	11.31 (13.81)	−11.46 (24.89)
Other touch	−92.70 (49.86)	55.62 (25.55)	−60.30 (38.76)

Table 7 For the twins data, estimations of γ obtained using the proposed multivariate response model with random effect. Standard errors (in brackets) are obtained via the bootstrap (1000 replicates).

	Multivariate response model		
	Depression	Stress	Anxiety
Self touch	−26.82 (36.30)	12.10 (13.15)	−7.12 (23.81)
Other touch	−83.43 (47.50)	46.82 (16.00)	−73.72 (29.35)

5.2 Import and export data

The considered data set provides country-wise percentages of imports and exports, measured in million USD, in relation to overall GDP, for 44 countries, between 2018 and 2022. A varying number of observations has been available for different countries during this time. Specifically, Australia, Japan, Korea, Mexico, New Zealand, Turkey, the United States, China and Colombia have four observations each, while India, Russia and Brazil have three observations each. The remaining countries have five observations each. The data are extracted from the OECD website (Organisation for Economic Co-operation and Development, 2023b) and are available as **trading_data** in R package mult.latent.reg. In our analysis, the logs of imports and exports constitute a bivarate response variable, with $r = 44$ countries defining the upper-level, and $n_i \in \{3, 4, 5\}, i = 1, \ldots, r$, repeated measurements on the lower-level.

Fitting a bivariate response model of type (2.1), but without covariate, the minimum AIC is attained for $K = 4$ mass points (Table 8). For each country, we obtain the posterior probabilities w_{ik} according to (3.15), an excerpt of the full matrix $(w_{ik})_{1 \leq i \leq r, 1 \leq k \leq K}$ is given in Table 9. The countries are ordered in this table by their posterior intercepts $z_i^* = \sum_{k=1}^{K} w_{ik} \hat{z}_k$ (Aitkin, 1996), with smaller values corresponding to smaller import/export volume relative to GDP. We can think of this column as representing predicted values of a latent variable which we could describe as 'international trade volume per GDP'. So, according to this linearized view on the problem, Luxembourg shows the largest trade volume per GDP, and the United State the smallest.

A sensible way of clustering the observations is to follow the MAP rule, that is, each upper-level unit (country) i is assigned to the cluster k to which it belongs with the largest probability w_{ik}. We can see from the last column in Table 9 that, according to this rule, Luxembourg is the only country assigned to its high-volume mass point. The second-largest mass point encompasses a wide range of countries ranging from Ireland to Germany, followed by the second smallest mass point featuring

Table 8 The values of AIC$=-2\ell + 2q$ for the trading data fitted with different number of mixture components. The best solution is highlighted in bold.

K	2	3	4	5	6
ℓ	−63.397	−52.227	−39.935	−42.939	−42.256
q	11	15	19	23	27
AIC	148.795	134.455	**117.870**	131.879	138.512

countries from Iceland to Israel. The mass point corresponding to the smallest trading volume per GDP comprises of 10 countries, most of which very large countries including Australia and all BRIC countries. Figure 4 top left provides a graphical representation of this clustering approach, with observations coloured by MAP classifications.

The second mass point ($k = 2$) has smaller variance compared to the first and third mass points (see Table 10), and all countries allocated to this cluster according to the MAP rule share some probability mass with the third cluster (but not all countries in the third cluster share probability mass with the second). There is no obvious characteristic distinguishing these clusters, even though countries in the third cluster tend to be smaller in size, especially those that have no or little probability mass shared with the second cluster.

We note that neither the ranking (by posterior intercepts) nor the clustering (by the MAP rule) gives clear evidence on how *well* two countries, or two clusters, can actually be distinguished. However, the posterior probabilities available in the inner part of Table 9 help us to provide a principled way of doing so. If the largest posterior probability of the observation exceeds a certain level of confidence, say 0.95, it is clustered into that specific cluster with 95% confidence. That is, it can be robustly distinguished from observations (countries) that are allocated to *other* mass points at this level of confidence. This enables us to produce a 'robust' clustering of countries, as illustrated in Figure 4 top right. For example, Luxembourg is classified to the highest mass point 4 with a probability of 1, and it can be reliably distinguished from countries such as Ireland, Slovak Republic, …, Austria that all have a posterior probability > 0.95 of belonging to mass point 3. Conversely, all countries for which the largest posterior probability is less than 0.95 are considered an uncertain observation that does not belong to any specific mass point, coloured as grey points in Figure 4 top right. This specifically concerns the countries of France, Turkey and the United Kingdom with their largest probabilities being below 0.95. At this level of confidence, the second mass point is eradicated entirely. However, changing the confidence level to 90% allows a robust clustering of these countries to mass point 2, as shown in Figure 4 bottom. Further conclusions could be drawn with careful reasoning: for instance, even under a 95% level of confidence, all countries from Greece to the United Kingdom can be robustly distinguished from both mass points 1 and 4, as they feature > 95% probability mass between them. They just cannot be robustly distinguished between mass points 2 and 3 at that level.

We have seen that different levels of confidence lead to different 'confidence-adaptive' allocations of observations to clusters. While the most appealing choices of the confidence level for this purpose appear to be 1 (certain allocation), 0.95 and 0.90 (as illustrated), values down to even 0.5 could be of interest in certain situations as they would ensure a more certain allocation than the MAP estimate. It is also noteworthy that, as we have seen by the example of cluster 4, which, according to Table 9, only consists of one element, the *size* of a subpopulation is not of relevance for it being robustly clustered.

Table 9 Classification and ranking for the trade and service data with $K = 4$. Posterior probabilities: ▨ $0.10 < p < 0.90$, ▨ $0.90 \leq p < 0.95$, ▨ $0.95 \leq p < 1$.

Country	posterior intercept		Mass points				MAP
		k	1	2	3	4	
		$\hat{\pi}_k$	0.236	0.310	0.431	0.023	
	z_i^*	$\hat{z}_k$	−1.402	−0.321	0.846	2.933	
United States	−1.402		1.000	0.000	0.000	0.000	1
Brazil	−1.401		1.000	0.000	0.000	0.000	1
Japan	−1.401		0.999	0.001	0.000	0.000	1
China	−1.401		0.999	0.001	0.000	0.000	1
India	−1.401		0.999	0.001	0.000	0.000	1
Colombia	−1.399		0.999	0.001	0.000	0.000	1
Indonesia	−1.378		0.979	0.021	0.000	0.000	1
Australia	−1.378		0.979	0.021	0.000	0.000	1
Russia	−1.373		0.974	0.025	0.001	0.000	1
New Zealand	−1.321		0.928	0.070	0.002	0.000	1
Israel	−0.574		0.260	0.715	0.025	0.000	2
South Africa	−0.383		0.099	0.862	0.039	0.000	2
Canada	−0.325		0.056	0.896	0.048	0.000	2
United Kingdom	−0.292		0.035	0.907	0.058	0.000	2
Turkey	−0.272		0.025	0.909	0.066	0.000	2
France	−0.251		0.018	0.906	0.076	0.000	2
Chile	−0.218		0.010	0.893	0.097	0.000	2
Italy	−0.211		0.009	0.889	0.102	0.000	2
Costa Rica	−0.156		0.004	0.850	0.146	0.000	2
Republic of Korea	−0.149		0.004	0.844	0.152	0.000	2
Spain	−0.127		0.003	0.828	0.169	0.000	2
Mexico	−0.080		0.002	0.790	0.208	0.000	2
Norway	−0.040		0.021	0.719	0.260	0.000	2
Finland	0.156		0.000	0.591	0.409	0.000	2
Iceland	0.180		0.000	0.570	0.430	0.000	2
Germany	0.357		0.000	0.418	0.582	0.000	3
Sweden	0.490		0.000	0.304	0.696	0.000	3
Portugal	0.492		0.000	0.303	0.697	0.000	3
Greece	0.614		0.000	0.198	0.802	0.000	3
Austria	0.790		0.000	0.047	0.953	0.000	3
Poland	0.806		0.000	0.034	0.966	0.000	3
Denmark	0.824		0.000	0.018	0.982	0.000	3
Switzerland	0.839		0.000	0.005	0.995	0.000	3
Latvia	0.843		0.000	0.002	0.998	0.000	3
Czech Republic	0.844		0.000	0.001	0.999	0.000	3
Estonia	0.845		0.000	0.000	1.000	0.000	3
Netherlands	0.845		0.000	0.000	1.000	0.000	3
⋮	⋮		⋮	⋮	⋮	⋮	⋮
Slovak Republic	0.846		0.000	0.000	1.000	0.000	3
Ireland	0.846		0.000	0.000	1.000	0.000	3
Luxembourg	2.933		0.000	0.000	0.000	1.000	4

Table 10 Estimated σ_{lk}, where $l = 1, 2$ and $k = 1, \ldots, 4$ for the fitted model in section 5.2.

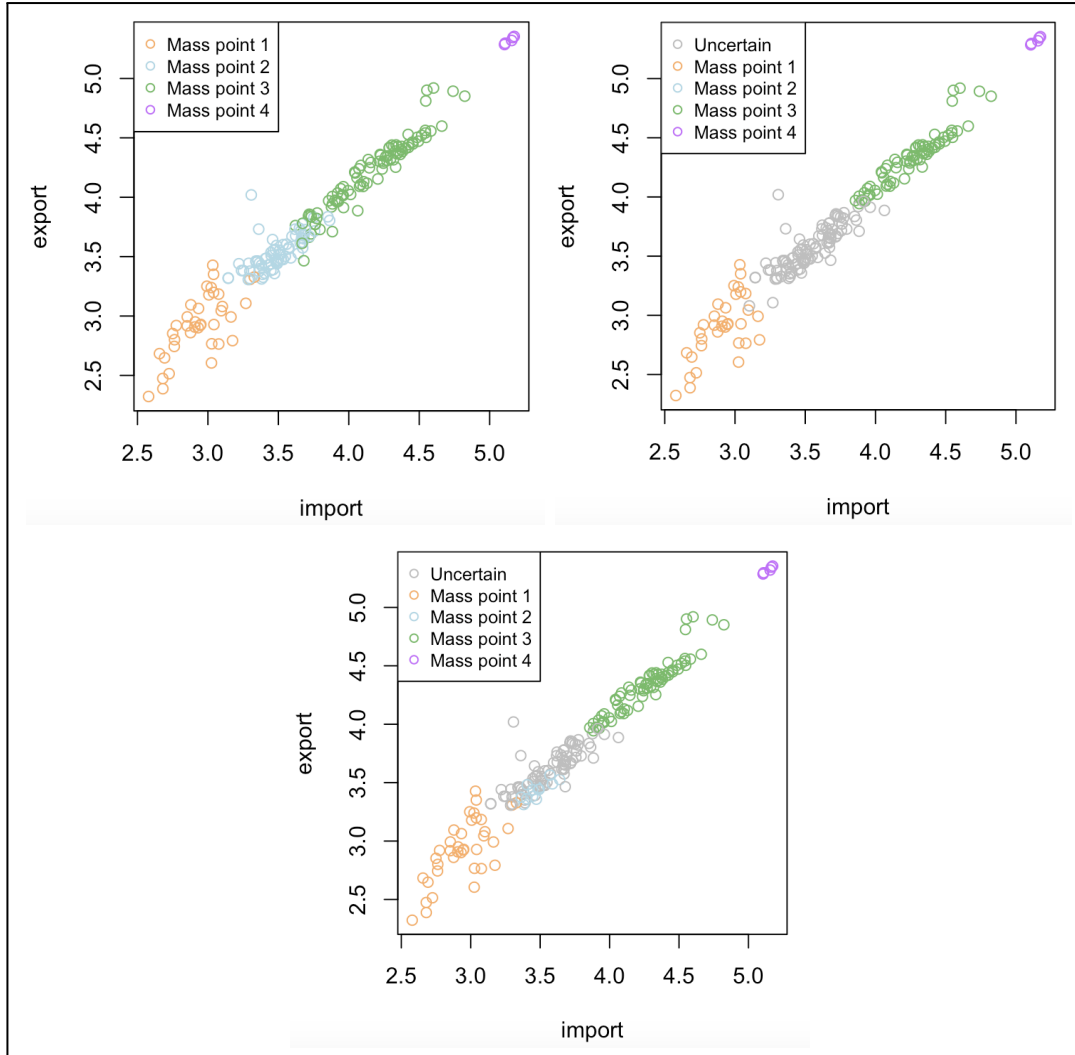|	Mass points			
k	1	2	3	4
$\hat{\sigma}_{1k}$	0.204	0.177	0.329	0.039
$\hat{\sigma}_{2k}$	0.285	0.172	0.347	0.036

Figure 4 Clustering of the imports and exports data; top left using the MAP rule; top right with 95% confidence; bottom with 90% confidence. Note that three to five observations correspond to each country.

Table 11 The values of AIC$= -2\ell + 2q$ for the PIAAC data fitted with different number of mixture components. The best solution is highlighted in bold.

K	2	3	4	5	6
ℓ	−342.243	−330.358	−324.851	−325.125	−325.330
q	21	26	31	36	41
AIC	726.485	712.716	**711.702**	722.251	732.660

Table 12 The estimates of covariate coefficients (matrix Γ) for the PIAAC data. Standard errors (in brackets) are obtained via the bootstrap.

	Gender	Employment status
Literacy	−0.417 (1.336)	2.683 (1.336)
Numeracy	8.817 (1.458)	−0.517 (1.427)
Problem solving	1.833 (1.163)	6.056 (1.218)

5.3 PIAAC survey of adult skills

We now analyse the PIAAC data set, where literacy, numeracy and problem solving constitute a three-variate response, and gender and employment status serve as two covariates. Again this is a two-level model, with 30 countries defining the upper-level. The lower-level is defined through the different combinations of the covariate factor levels within each country; that is, there are four lower-level 'observations' for each country corresponding to the average score for this combination of covariates. More details about the survey and its design are provided in part D of the supplementary material and on the OECD website (Organisation for Economic Co-operation and Development, 2023a). It important to point out at this point that, in the analysis provided in here, no adjustment for country-wise variations in sampling design was undertaken, unlike for instance in the work by Hämäläinen et al. (2017) who fitted survey-weighted logistic models to PIAAC problem-solving scores.

Using the methodology devised in this article, an examination of AIC values across different values of K (Table 11) shows that a minimum AIC value of 711.702 is attained for $K = 4$; hence we use this choice of K to fit model (2.1). Posterior intercepts can again be obtained through the use of $z_i^* = \sum_{k=1}^{K} w_{ik} \hat{z}_k$, with posterior probabilities w_{ik} according to (3.15). These posterior intercepts can be seen as the summary information for each country, providing the residual performance after the covariates have been taken into account. The role of the covariates is to 'take out' the effects of such variables in the clustering process. The estimates of the covariate coefficients are shown in Table 12. The results show how gender (male = 1, female = 0) and employment status (employee = 1, self-employed = 0) relate to literacy, numeracy, and problem-solving skills. For example, it indicates that employees have expected problem-solving scores that are 6.056 higher than for self-employed. For literacy, the advantage of the employees reduces to 2.683 units, while for numeracy, the self-employed tend to fare better by 0.517 units. Providing the z_i^* in rank order results in a league table, shown in Table 13. The posterior probabilities obtained at the convergence of the EM algorithm are also given in this table, and can be used for classification of countries according to their skill levels.

We can distinguish two countries in terms of their cluster membership if they fall with 95% confidence into two different mass points. In Table 13, all countries from New Zealand to Hungary

Table 13 Classification and ranking for the PIAAC data using model $x_{ij} = \alpha + \beta z_i + \Gamma v_{ij} + \varepsilon_{ij}$ with $K = 4$. Posterior probabilities: □ $0.05 < p < 0.10$, ▦ $0.10 < p < 0.90$, ▨ $0.90 \leq p < 0.95$, ▰ $0.95 \leq p < 1$.

Country	Posterior intercept	Mass points			
		0.100 −2.650	0.115 −0.634	0.275 −0.069	0.510 0.700
Chile	−2.650	1.000	0.000	0.000	0.000
Mexico	−2.650	1.000	0.000	0.000	0.000
Turkey	−2.650	1.000	0.000	0.000	0.000
Greece	−0.632	0.000	0.997	0.003	0.000
Spain	−0.603	0.000	0.945	0.055	0.000
Republic of Korea	−0.592	0.000	0.926	0.074	0.000
Italy	−0.296	0.000	0.403	0.597	0.000
United States	−0.100	0.000	0.066	0.927	0.007
Poland	−0.092	0.000	0.058	0.930	0.012
Slovenia	−0.065	0.000	0.018	0.963	0.019
Ireland	−0.010	0.000	0.013	0.902	0.085
France	−0.006	0.000	0.007	0.905	0.088
Israel	0.012	0.000	0.005	0.885	0.110
England (UK)	0.023	0.000	0.022	0.843	0.135
Denmark	0.398	0.000	0.000	0.392	0.608
Germany	0.451	0.000	0.000	0.323	0.667
Flanders (Belgium)	0.567	0.000	0.000	0.173	0.827
Norway	0.654	0.000	0.000	0.060	0.940
Czech Republic	0.659	0.000	0.000	0.054	0.946
Hungary	0.667	0.000	0.000	0.043	0.957
Austria	0.676	0.000	0.000	0.031	0.969
Australia	0.683	0.000	0.000	0.021	0.979
Estonia	0.686	0.000	0.000	0.018	0.982
Finland	0.689	0.000	0.000	0.014	0.986
Canada	0.692	0.000	0.000	0.010	0.990
Japan	0.696	0.000	0.000	0.005	0.995
Slovak Republic	0.696	0.000	0.000	0.005	0.995
Netherlands	0.699	0.000	0.000	0.002	0.998
Sweden	0.699	0.000	0.000	0.001	0.999
New Zealand	0.700	0.000	0.000	0.000	1.000

are assigned with $> 95\%$ confidence to the best-performing mass point 4. As such, they can be robustly distinguished from Slovenia and all countries with lower posterior intercepts, which all have posterior probabilities of at least 95% of pertaining to any of the first three mass points. Slovenia is the only country which is allocated with at least 95% confidence to mass point 3; hence, it can be robustly concluded to have performed better than Greece, Turkey, Mexico and Chile. Greece, in turn, is the only country which is clustered to mass point 2 with more than 95% confidence and, as such, can be robustly distinguished from the three countries belonging to the worst-performing mass point 1.

Note that it is not possible to determine a comparative ranking among countries belonging to the same mass point, for example, we cannot say that Japan has performed better than Canada. We also cannot robustly conclude that England (UK) or even Flanders (Belgium) have performed better than the United States, since they all share at least 5% probability mass with mass point 3.

6 Concluding remarks

We have provided a novel methodological approach for the inclusion of a random effect into multivariate response models, based on the NPML method for mixture models. The proposed approach enables us to accurately estimate covariate effects under the presence of correlations between response variables. Crucially, such correlations impact the standard errors of parameter estimates, as observed in our simulation studies and real data applications. It should however be noted that under this methodology, no analytic calculation of the standard errors is possible, hence requiring us to resort to bootstrap techniques.

Another advantage of the proposed methodology is in providing the matrix of posterior probabilities produced alongside the estimation process, as well as in calculating posterior random effects, based on the fitted model. We have demonstrated how these can be used for model-based clustering along the direction of the latent subspace and conditional on covariate values. The clustering can be performed either directly based on the maximum a posteriori (MAP) rule or can be driven by a user-specified degree of confidence in the cluster allocation, allowing for fine-grained insights into the separability of upper-level units on the scale of their posterior random effect. As suggested by a referee, entropy measures could potentially be used to assess the uncertainty of posterior probabilities, which in our context would correspond to values $-\sum_{k=1}^{K} w_{ik} \log w_{ik}$, with high values indicating high uncertainty. Some care is needed with this approach, as many of the posterior probabilities in our context are zero.

Computationally, the proposed method can be regarded as the multivariate extension of the `allvc` function available in R package npmlreg (Einbeck et al., 2018). We here focused on the Gaussian errors assumption for the response model and used the nonparametric maximum likelihood approach to handle the marginal density of x_{ij}. In contrast, the `allvc` function is based on the glm framework, hence allowing any arbitrary exponential family distribution for the response. The extension of the proposed work to an exponential family framework is not straightforward and requires further research. A possible starting point for such research is to consider methods from the 'homogeneity pursuit' literature such as the pairwise composite likelihood in Hui et al. (2018). Furthermore, the `allvc` function supports Gaussian quadrature besides a nonparametric maximum likelihood variant, which would also appear to be feasible for our scenario even though we have not been able to identify a strong incentive to do so. As a further limitation of the current state of development, one could mention that we are not yet able to accommodate weights, which would have been useful for the PIAAC data to adjust for country-wise differences in sampling design.

Acknowledgements

for our analysis in Section 5.1, and Dr Nick Sofroniou for helpful suggestions in the preparation of this project. We would also like to thank the three reviewers for their constructive and insightful comments.

Supplementary material

Supplementary material for this article is available online.

References

Aitkin M (1996)   Empirical Bayes shrinkage using posterior random effect means from nonparametric maximum likelihood estimation in general random effect models. In *Proceedings of the 11th International Workshop on Statistical Modelling*, pages 87–94.

Aitkin M (1999) A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, **55**, 117–28.

Bartolucci F, Pennoni F and Vittadini G (2011) Assessment of school performance through a multilevel latent Markov Rasch model. *Journal of Educational and Behavioral Statistics*, **36**, 491–522.

Bates D, Mächler M, Bolker B and Walker S (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48. doi: 10.18637/jss.v067.i01.

Dempster AP, Laird NM and Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**, 1–22.

Di Mari R, Bakk Z, Oser J, and Kuha J (2023) A two-step estimator for multilevel latent class analysis with covariates. *Psychometrika*, **88**, 1144–1170. doi: 10.1214/16-AOAS988.

Einbeck J Darnell R and Hinde J (2018) *npmlreg: Nonparametric maximum likelihood estimation for random effect models*. URL https://CRAN.R-project.org/package= npmlreg. R package version 0.46-5.

Gnaldi M, Bacci S, and Bartolucci F (2016) A multilevel finite mixture item response model to cluster examinees and schools. *Advances in Data Analysis and Classification*, **10**, 53–70.

Goodman LA (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**, 215–31.

Grilli L, Pennoni F, Rampichini C and Romeo I (2016)   Exploiting TIMSS and PIRLS combined data: multivariate multilevel

modelling of student achievement. *The Annals of Applied Statistics*, **10**, 2405–26.

Hui FK, Müller S and Welsh A (2018) Sparse pairwise likelihood estimation for multivariate longitudinal mixed models. *Journal of the American Statistical Association*, **113**, 1759–69.

Hämäläinen R, Wever BD, Nissinen K and Cincinnato S (2017) Understanding adults' strong problem-solving skills based on piaac. *Journal of Workplace Learning*, **29**, 537–53.

Lukočienė O, Varriale R and Vermunt JK (2010) The simultaneous decision (s) about the number of lower-and higher-level classes in multilevel latent class analysis. *Sociological Methodology*, **40**, 247–83.

Marques da Silva Júnior AH, Einbeck J and Craig PS (2018) Fisher information under Gaussian quadrature models. *Statistica Neerlandica*, **72**, 74–89.

Masci C, Ieva F and Paganoni AM (2022) Semiparametric multinomial mixed-effects models: A university students profiling tool. *Annals of Applied Statistics*, **16**, 1608 –32.

Meng X-L and Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**, 267–78.

Organisation for Economic Co-operation and Development (2023a) Main elements of the Survey of Adult Skills. https://www.oecd.org/skills/piaac/mainelementsofthesurveyofadultskills.htm. Accessed on 2023-05-29.

Organisation for Economic Co-operation and Development (2023b) Trade in goods and services. https://data.oecd.org/trade/trade-in-goods-and-services.htm. Accessed on 2023-05-29.

Pennoni F, Bartolucci F and Pandolfi S (2024) Variable selection for hidden Markov models with continuous variables and missing data. *Journal of Classification*, **41**, 1–22.

Reissland N, Einbeck J, Wood R and Lane A (2021) Effects of maternal mental health on prenatal movement profiles in twins and singletons. *Acta Paediatrica*, **110**, 2553–58.

Verbeke G, Fieuws S, Molenberghs G and Davidian M (2014) The analysis of multivariate longitudinal data: a review. *Statistical Methods in Medical Research*, **23**, 42–59.

Vermunt JK (2003) Multilevel latent class models. *Sociological Methodology*, **33**, 213–39.

Vermunt JK (2008) Multilevel latent variable modelling: An application in education testing. *Austrian Journal of Statistics*, **34**, 285–99.

Zhang Y and Einbeck J (2024a) A versatile model for clustered and highly correlated multivariate data. *Journal of Statistical Theory and Practice*, **18**. doi: 10.1007/s42519-023-00357-0.

Zhang Y and Einbeck J (2024b) *mult.latent.reg: Regression and clustering in multivariate response scenarios*. URL https://CRAN.R-project.org/package=mult.latent.reg. R package version 0.2.1.

Zhang Y and Einbeck J (2024c) R package mult.latent.reg for multivariate response scenarios with latent structures. In *Proceedings of the 38th International Workshop on Statistical Modelling*, pages 320–23.