Journal of Statistical Theory and Practice (2024) 18:5
https://doi.org/10.1007/s42519-023-00357-0

ORIGINAL ARTICLE

A Versatile Model for Clustered and Highly Correlated Multivariate Data

Yingjuan Zhang[1] · Jochen Einbeck[1,2]

Accepted: 17 November 2023

Abstract
For the analysis of multivariate data with an approximately one-dimensional latent structure, it is suggested to model this latent variable by a random effect, allowing for the use of mixed model methodology for dimension reduction purposes. We implement this idea through the mixture-based approach for the estimation of random effect models, hence conveniently enabling clustering of observations along the latent linear subspace, and derive the estimators required for the ensuing EM algorithm under several error variance parameterizations. A simulation study is conducted, and several important inferential problems, including clustering, projection, ranking, regression on covariates, and regression of an external response on the predicted latent variable, are considered and illustrated by real data examples.

Keywords Linear approximation · Nonparametric maximum likelihood · MAP estimation · Bootstrap · Multivariate response models

1 Introduction

It is not uncommon for a set of variables to be so strongly correlated that they can be considered as intrinsically one-dimensional, meaning that they can be considered to be generated by some latent one-dimensional linear subspace plus noise. As examples for such situations, one could name price indexes for several goods, or educational attainment scores on several different abilities, or several psychological mental health indicators. While a rich set of statistical tools exists for identifying best linear approximations of multivariate data, usually based on algebraic properties of the sample

✉ Yingjuan Zhang
yingjuan.zhang@durham.ac.uk

Jochen Einbeck
jochen.einbeck@durham.ac.uk

[1] Department of Mathematical Sciences, Durham University, Durham, UK

[2] Durham Research Methods Centre, Durham, UK

Published online: 03 January 2024

covariance matrix (such as principal component analysis), a different approach is followed in this paper which is firmly rooted in basic principles of statistical modelling, and hence allows versatile access to routine statistical tasks such as clustering or regression.

The basic idea, which is of more general validity than the framework focused on in this paper, is to consider the approximating lower-dimensional subspace as a latent variable in a multivariate statistical model and to model this latent variable by a random effect. In this work, we develop and implement this very general idea in a more specific framework, where we assume the low-dimensional structure to be a one-dimensional space, i.e. a straight line.

More specifically, we consider a scenario where the multivariate data $x_i \in \mathbb{R}^m$ are noisy observations scattered along the one-dimensional space $\alpha + \beta z$, where $\alpha, \beta \in \mathbb{R}^m$, and $z \in \mathbb{R}$ is an unobserved instance of a (latent) variable Z. Then, the observed data $x_i = (x_{i1}, \ldots, x_{im})^T, i = 1, \ldots, n$, are assumed to be generated from the following random effect model,

$$x_i = \alpha + \beta z_i + \varepsilon_i, \tag{1}$$

where $\varepsilon_i \sim N(0, \Sigma_i)$ is m-variate Gaussian noise with a positive definite variance matrix $\Sigma_i \equiv \Sigma(z_i) \in \mathbb{R}^{m \times m}$. It is clear that a model with observation specific $m \times m$ variance matrices is heavily overparametrized, and we will never contemplate fitting this model in full generality. We still provide this general notation in Eq. (1) as it contains all practically relevant special cases that will be naturally arising, including, of course, the homoscedastic case $\Sigma_i = \Sigma, i = 1, \ldots, n$.

For the estimation of the random effect distribution along this line, we use the nonparametric maximum likelihood approach, which amounts to representing this distribution by a set of discrete mass points (mixture centres) with some corresponding masses (mixture probabilities). While this may look like a restrictive assumption, it is actually more flexible than the application of a Gaussian random effect, as it allows for multi-modalities in the distributions of the latent variable. Indeed, the mixture character of this approach allows for clustering of observations based on the fitted model.

In consequence, this arrives at a modelling approach with an enormous versatility. Firstly, as just expressed, observations can be clustered based on maximum a posteriori (MAP) probabilities of class membership [14, chapter 11]. Secondly, projecting the original data points onto the estimated lower-dimensional space, the dimension of the original multivariate data is reduced (to 1, in the simple framework as discussed in this work), and the compressed data can be used as summary statistic (such as an overall price index across several goods) or for further inferential purposes. Thirdly, the relative order of the posterior random effects (observations 'projected' onto the latent linear subspace) can be used for ranking observations in multivariate data sets. Finally, we will show that it is not difficult to include additional covariates into model (1) so that one has de facto a novel tool for multivariate response situations, yielding reduced parameter standard errors as compared to the separate univariate response models. We will give each of these important applications some prominence later in the paper.

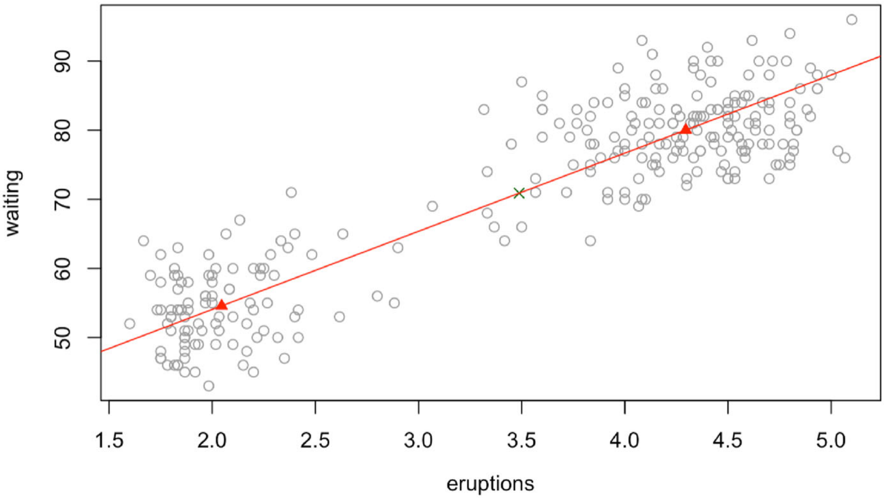

Fig. 1 Graph showing the estimated one-dimensional space with cluster centres in red and α in green

To enable some intuition for how the model operates (in the simpler case without covariates), we use the faithful data set in R package **MASS** [20]. This is a two-dimensional data set with 272 observations and two variables: eruption time and the waiting time between two eruptions. The straight line in Fig. 1 is the one-dimensional latent space $\alpha + \beta z$ that is parameterized by the latent variable. The red triangles positioned along the straight line are the estimated mixture (cluster) centres. To give some metaphor, one could consider the mixture centres as 'washing pegs' spanning a 'washing rope' holding the clusters. We will return to this example in Sect. 5.1 and illustrate there in detail how exactly this image translates into projections (dimension reduction) and clusterings.

Some methodologically related techniques have been previously suggested in the literature, partly very long ago. In the homoscedastic case, the model (1) can be seen as a one-dimensional factor analysis model (see [14, chapter 12]), with the difference that we will apply a discrete mixture approximation of the latent variable z_i. There is also some overlap with the generative topographic mapping (GTM, [7]), which allows for nonlinear manifolds rather than just a latent straight line. However, in the GTM, the latent variables are parameterized by a fixed and equidistant grid, rather than estimable masses and mass points, rendering the approach less useful for clustering-type applications. Under both the factor analysis and the GTM approaches, there is no immediate possibility to include covariates, and hence, they do not serve as a multivariate response model. Models of the type (1) have also been proposed in the literature on model-based clustering in high-dimensional data scenarios, an overview over which has been given in [8]. Sammel et al. [17] proposed a latent variable model for mixed discrete and continuous outcomes from the exponential family, where, however, the latent variable itself is modelled by covariates, contrasting with the approach investigated in here.

The remainder of this work is organized as follows. In Sect. 2, we give details of the nonparametric maximum likelihood procedure to estimate the parameters of model (1), yielding an EM algorithm which also automatically estimates masses, mass points, and posterior probabilities of data points being associated with those. Simulation studies which illustrate the accuracy of the proposed estimation methodology are presented in Sect. 3. This is followed by Sect. 4, where we will lay down the clustering and projection operations explicitly. Furthermore, in Sects. 4.4 and 4.5, we consider extension of the proposed framework allowing for covariates along with a bootstrap approach for the computation of standard errors. Applications to several real data scenarios are given in Sect. 5, which we also use to illustrate the main application pillars of clustering, dimension reduction, ranking, and regression, explicitly. The paper is concluded with a discussion in Sect. 6. Some technical derivations are relegated to an 'Appendix'.

2 Methods and Estimation

2.1 Likelihood

The marginal probability density function $f(x_i|\alpha, \beta)$ for observations generated from model (1) can be written as

$$f(x_i|\alpha, \beta) = \int f(x_i, z_i|\alpha, \beta)\mathrm{d}z_i = \int f(x_i|z_i, \alpha, \beta)\phi(z_i)\mathrm{d}z_i, \tag{2}$$

where $f(x_i, z_i|\alpha, \beta)$ is the joint probability distribution of observed data x_i and unobserved random effects z_i, and $\phi(\cdot)$ is the density function of the random effect distribution Z. This model is not fully specified since it lacks specific parametrizations of the (unknown) $\Sigma_i = \Sigma(z_i) = \mathrm{Var}(x_i|z_i, \alpha, \beta)$ and ϕ, but let us consider any (additional) parameters involved into these initially as nuisance parameters and construct appropriate parametrizations for these as we go along.

The initial goal is to find maximum likelihood estimates for the parameters α and β in model (1). Building on the marginal density (2), the likelihood of model (1) is the following,

$$L(\alpha, \beta) = \prod_{i=1}^{n} \int f(x_i|z_i, \alpha, \beta)\phi(z_i)\mathrm{d}z_i$$

with corresponding log-likelihood,

$$l(\alpha, \beta) = \sum_{i=1}^{n} \log\left\{\int f(x_i|z_i, \alpha, \beta)\phi(z_i)\mathrm{d}z_i\right\}. \tag{3}$$

At this stage, a decision needs to be made on how to deal with the integral figuring in Eq. (3). In principle, one could do this based on a Gaussianity assumption on $\phi(\cdot)$, as common in the mixed model context, in this case leading us back to a factor analysis framework. However, for reasons expressed in the introduction, we have decided here differently and employ instead Aitkin's nonparametric maximum likelihood approach [2]. Here, the random effect distribution Z is approximated by a discrete mixture distribution, say $\tilde{Z}$, which is supported on a finite number of mass points $z_1, \dots, z_K$ with masses $P(\tilde{Z} = z_k) = \pi_k, k = 1, \dots, K$. This discrete mixture facilitates a simple approximation of the marginal likelihood which just involves sums rather integrals, i.e.

$$l(\alpha, \beta) \approx \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} f(x_i | z_k, \alpha, \beta) \pi_k \right\}. \tag{4}$$

Laird [12] showed that the marginal likelihood (3) can be approximated arbitrarily well by (4) with a finite set of mass points. We see that this has now become a mixture-type problem, with each mixture component k representing a latent 'class' within the domain of Z (we will use the terms class and component interchangeably henceforth). The EM algorithm [9] is one of the most widely used algorithms for the estimation of parameters in mixture models.

Denote by $f_{ik} = P(x_i | \tilde{Z} = z_k) = f(x_i | z_k, \alpha, \beta)$ the probability density of x_i conditional on class k. Then, we know that

$$P(x_i, \tilde{Z} = z_k) = P(x_i | \tilde{Z} = z_k) P(\tilde{Z} = z_k) = f_{ik} \pi_k.$$

Since it is in practice unknown which component each observations belongs to, this is an incomplete data scenario. We describe the missing information on the component membership by an indicator variable

$$G_{ik} = \begin{cases} 1, & \text{if observation } i \text{ belongs to component } k \\ 0, & \text{otherwise.} \end{cases}$$

This defines complete data $(x_i, G_{i1}, \dots, G_{iK})$, $i = 1, \dots, n$, with probability

$$P(x_i, G_{i1}, \dots, G_{iK}) = \prod_{k=1}^{K} (f_{ik} \pi_k)^{G_{ik}}$$

and resulting complete data likelihood $\prod_{i=1}^{n} \prod_{k=1}^{K} (f_{ik} \pi_k)^{G_{ik}}$. Then, we can obtain the expected complete log-likelihood

$$\begin{aligned} l_c &= \sum_{i=1}^{n} \mathbb{E}\left[\log\left(\prod_{k=1}^{K} (\pi_k f_{ik})^{G_{ik}}\right) |x_i\right] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}\left[G_{ik}|x_i\right] \log\left(\pi_k f_{ik}\right) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik} \log \pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik} \log f_{ik} \end{aligned} \tag{5}$$

where $w_{ik} = \mathbb{E}\left[G_{ik}|x_i\right] = P(G_{ik} = 1|x_i) = P(\tilde{Z} = z_k|x_i)$, which is the probability of each observation i belonging to component k. For the component-specific densities f_{ik}, we specify, conditional on the mixture centres z_k, a multivariate Gaussian model

$$f_{ik} = \frac{1}{(2\pi)^{m/2}} \frac{1}{|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x_i - \alpha - \beta z_k)^T \Sigma_k^{-1} (x_i - \alpha - \beta z_k)\right) \tag{6}$$

where we allow the variance matrices $\Sigma_k = \Sigma(z_k)$ to depend on the cluster k but not on observation i, hence reducing the complexity of the original, fully heteroscedastic, variance specification considerably. The terms $\alpha+\beta z_k$ can be interpreted as the mixture centres in the original data space, spanned along the line $\alpha + \beta z$. Note that the right hand side of (4) is then the likelihood corresponding to the 'approximative' model

$$x_i|z_k, \alpha, \beta \sim N(\alpha + \beta z_k, \Sigma_k) \text{ with probability } \pi_k, \tag{7}$$

where we treat the mass points z_k, $k = 1, \ldots, K$, and their associated masses π_k as unknown parameters to be estimated in the EM algorithm alongside with the model parameters α and β. This model can be seen as a Gaussian mixture model with structured mean function and component-specific variances, or as a multivariate response version of the 'nonparametric maximum likelihood' (NPML) approach for the estimation of mixture masses and mass points in random effect models [2], [4, chapter 8].

Several reduced, parsimonious, parameterizations of the variance matrices Σ_k are possible in order to describe the shape of the clusters around the mixture centres. The simplest case (i) would be a constant and diagonal matrix $\Sigma_k \equiv \Sigma = \text{diag}(\sigma_j^2)_{\{1\le j\le m\}} \in \mathbb{R}^{m\times m}$, which gives the same variance specification to all K components of the mixture. Second (ii), we consider using different diagonal variance matrices for different components, $\Sigma_k = \text{diag}(\sigma_{jk}^2)_{\{1\le j\le m\}} \in \mathbb{R}^{m\times m}$, which yields an improvement for estimating data that have clusters of different sizes. Third (iii), we consider using the same full (unrestricted) variance matrix, $\Sigma_k \equiv \Sigma \in \mathbb{R}^{m\times m}$, to capture the correlation of variables. Finally (iv), different full (unrestricted) variance matrices, $\Sigma_k \in \mathbb{R}^{m\times m}$ give better estimations when dealing with clusters that differ by shape and size. In line with (6) and (7), our notation in what follows will be tailored to this most general case (iv); with the results for the reduced parameterizations naturally deriving from this.

2.2 EM Algorithm

Now we can set up the EM algorithm for estimating model (7). It is noted that the developments in this subsection are for a fixed number of components, K. The question of choosing K is considered as a model selection problem and will be addressed through the use of model selection criteria as illustrated in later sections.

E-step

Using the Bayes' theorem, we obtain the posterior probability of observation i belonging to component k,

$$w_{ik} = P(\tilde{Z} = z_k | x_i) = \frac{P(\tilde{Z} = z_k) P(x_i | \tilde{Z} = z_k)}{\sum_l P(\tilde{Z} = z_l) P(x_i | \tilde{Z} = z_l)} = \frac{\pi_k f_{ik}}{\sum_l \pi_l f_{il}}. \tag{8}$$

M-step

Using expression (6) for the component-wise densities f_{ik}, the expected complete data log-likelihood becomes

$$\begin{aligned} l_c = & \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \log(\pi_k) + \sum_{i=1}^{n} \sum_{k=1}^{K} -\frac{1}{2} w_{ik} \log(|\Sigma_k|) + \sum_{i=1}^{n} \sum_{k=1}^{K} -\frac{m}{2} \log(2\pi) w_{ik} \\ & + \sum_{i=1}^{n} \sum_{k=1}^{K} -\frac{1}{2} w_{ik} (x_i - \alpha - \beta z_k)^T \Sigma_k^{-1} (x_i - \alpha - \beta z_k). \end{aligned} \tag{9}$$

Taking partial derivatives of l_c with respect to each parameter gives the score equations. We then obtain the following estimators for the parameters α, β, z_k and π_k by setting these score equations to zeros and solving them:

$$\hat{\alpha} = \left(\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \hat{\Sigma}_k^{-1} \right)^{-1} \left(\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \hat{\Sigma}_k^{-1} (x_i - \hat{\beta} \hat{z}_k) \right) \tag{10}$$

$$\hat{\beta} = \left(\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \hat{\Sigma}_k^{-1} \hat{z}_k^2 \right)^{-1} \left(\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \hat{\Sigma}_k^{-1} (x_i - \hat{\alpha}) \hat{z}_k \right) \tag{11}$$

$$\hat{z}_k = \frac{\sum_{i=1}^{n} w_{ik} \hat{\beta}^T \hat{\Sigma}_k^{-1} (x_i - \hat{\alpha})}{\sum_{i=1}^{n} w_{ik} \hat{\beta}^T \hat{\Sigma}_k^{-1} \hat{\beta}}. \tag{12}$$

For the mixture probabilities, since $\sum_{k=1}^{K} \pi_k = 1$, we need to apply a Lagrange multiplier by letting $\partial \left(l - \lambda (\sum_{k=1}^{K} \pi_k - 1) \right) / \partial \pi_k = 0$, yielding

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} w_{ik}. \tag{13}$$

Estimators for the flexible variance specifications are again obtained by equating the corresponding partial derivatives to zero, giving results as follows:

(i) $\Sigma = \text{diag}(\sigma_j^2)_{\{1\leq j\leq m\}} \in \mathbb{R}^{m\times m}, k = 1, \ldots, K$,

$$\hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_{ij} - \hat{\alpha}_j - \hat{\beta}_j\hat{z}_k)^2; \tag{14}$$

(ii) $\Sigma_k = \text{diag}(\sigma_{jk}^2)_{\{1\leq j\leq m\}} \in \mathbb{R}^{m\times m}, k = 1, \ldots, K$,

$$\hat{\sigma}_{jk}^2 = \frac{\sum_{i=1}^{n} w_{ik}(x_{ij} - \hat{\alpha}_j - \hat{\beta}_j\hat{z}_k)^2}{\sum_{i=1}^{n} w_{ik}}; \tag{15}$$

(iii) $\Sigma = \Sigma_1 = \cdots = \Sigma_k \in \mathbb{R}^{m\times m}, k = 1, \ldots, K$,

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k)(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k)^T; \tag{16}$$

(iv) $\Sigma_k \in \mathbb{R}^{m\times m}, k = 1, \ldots, K$,

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{n} w_{ik}(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k)(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k)^T}{\sum_{i=1}^{n} w_{ik}}. \tag{17}$$

It is evident that all of these estimators depend on the weights w_{ik}, hence requiring the use of the EM algorithm which iterates between finding the above estimates and updating the weights given the estimates.

2.3 Computational Considerations

It is noted from Eqs. (10), (11), and (12) that these involve many inversions of the estimated variance matrices $\hat{\Sigma}_k$. This can make the EM algorithm computationally unstable especiallly under the component-specific variance parameterizations (ii) and (iv). Therefore, in our implementation of the above EM algorithm, we disentangle the M-step updates of $\hat{\Sigma}_k$ from those of $\hat{\alpha}$, $\hat{\beta}$ and $\hat{z}_k$. Specifically, the updates (10), (11), and (12) are executed in a simplified form where $\hat{\Sigma}_k \equiv \text{diag}(\sigma^2)$, for some constant σ^2 which does not need to be specified since it cancels out from the resulting simplified update equations, yielding

$$\hat{\alpha} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i - \hat{\beta}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k\right); \tag{18}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k x_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k)}{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k^2 - \frac{1}{n}(\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\hat{z}_k)^2}; \tag{19}$$

$$\hat{z}_k = \frac{\hat{\beta}^T \sum_{i=1}^n w_{ik}(x_i - \hat{\alpha})}{\hat{\beta}^T \hat{\beta} \sum_{i=1}^n w_{ik}}. \tag{20}$$

That is, in our implementation, within each M-step, we cycle a small number times (five will be sufficient) between (18), (19), and (20), then we update $\hat{\pi}_k$ via (13), followed by the respective update of the variance matrices according to any of (14), (15), (16), or (17) depending on the variance parameterization. The resulting updated parameters are then used in the upcoming E-step (8). The simulation studies in Sect. 3 will confirm that this approach yields accurate parameter estimates.

2.4 Identifiability

Consider again the model for the x_i implied by equation (7), i.e.

$$x_i = \alpha + \beta z_k + \varepsilon_i. \tag{21}$$

The product term βz_k makes the parameters $\beta = (\beta_1, \ldots, \beta_m)^T$ and z_k unidentifiable. The vector α is also unidentifiable as, when moving along the estimated straight line, the same model could be attained by translating all z_k's along the line. Therefore, the model is identifiable only under certain restrictions, and in order to fix the problem, we standardize z_k by letting

$$\mathrm{E}(\tilde{Z}) = \sum_{k=1}^K \pi_k z_k = 0 \tag{22}$$

and

$$\mathrm{Var}(\tilde{Z}) = \sum_{k=1}^K \pi_k z_k^2 - (\pi_k z_k)^2 = 1. \tag{23}$$

Equation (22) solves the problem for α by fixing the position of z_k's along the estimated lower-dimensional subspace, and Eq. (23) solves the scale problem for β. Additionally, to identify the direction of the latent variable, we enforce $\beta_1 \geq 0$ (but any other component of β could equally be chosen for this).

2.5 Starting Values

Starting values can heavily influence the ability of the EM algorithm to locate the maximum of the likelihood (see, e.g. [15]). In the R implementation of the EM algorithm of our methodology, the following are the default starting values for parameters π_k, z_k, α, β, and Σ_k:

$$\pi_k^{(0)} = \frac{1}{K},$$

where K is the number of components. We use random numbers from a standard normal distribution as the starting values for the mass points,

$$z_k^{(0)} \sim N(0, 1),$$

which are then re-scaled according to (22) and (23). As default starting values for the line parameters, we use

$$\alpha^{(0)} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$\beta^{(0)} = x_r - \alpha^{(0)},$$

where $x_r \in \mathbb{R}^m$ is a randomly selected observation. For all four variance parameterizations, we use a diagonal matrix $\Sigma^{(0)} \in \mathbb{R}^{m\times m}$, not depending on k, as the 'starting variance matrix'. Let

$$s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2},$$

where $j = 1, 2, \dots, m$ and $\bar{x}_j$ is the sample mean of the j-th variable. Then, for each diagonal element $(\sigma_j^{(0)})^2$ of the diagonal matrix $\Sigma^{(0)}$, one has the starting value

$$\sigma_j^{(0)} = \frac{s_j}{K}, \quad j = 1, \dots, m.$$

3 Simulation

3.1 Estimation of Model Parameters

The EM algorithm derived in the previous section, with all four variance parameterizations, is implemented in R. Some simulations are set up to test the accuracy of the R implementation under different settings.

Under the variance parameterization (i), i.e. the same diagonal matrix for all components, we use two-dimensional data with three individual sample sizes $n = 100$, $n = 300$, and $n = 500$ and generate 1000 data sets from model (7) for each sample size. The true parameter values used for the simulations can be read from the first column of Table 1.

The methodology from Sect. 2.2 is then applied on each generated data set (with random starting values according to Sect. 2.5 to initialize the EM algorithm), and the 1000 estimates for each model parameter (see Table 3) are collected. Comparing the average of the estimated values to the true values of the parameters used to generate these data, some key results are shown in Table 1, Figs. 2, 3, and 4. In Table 1, the averaged estimates of the parameters are close to their true values across all parameters

Table 1 Simulation results under variance parameterization (i)

	True	Average estimates		
		$n = 100$	$n = 300$	$n = 500$
π_1	0.0500	0.0463	0.0507	0.0498
π_2	0.2500	0.2518	0.2504	0.2512
π_3	0.7000	0.7019	0.6988	0.6990
z_1	−0.6171	−0.6186	−0.6193	−0.6191
z_2	1.1675	1.2262	1.1693	1.1708
z_3	2.8023	2.8457	2.8130	2.8119
α_1	−1.000	−0.9936	−0.9985	−0.9985
α_2	1.000	1.0235	1.0036	0.9982
β_1	1.000	0.9915	0.9986	0.9966
β_2	3.000	2.9974	2.9982	2.9899
σ_1	0.5000	0.5043	0.4966	0.4985
σ_2	2.0000	1.9866	1.9892	1.9912

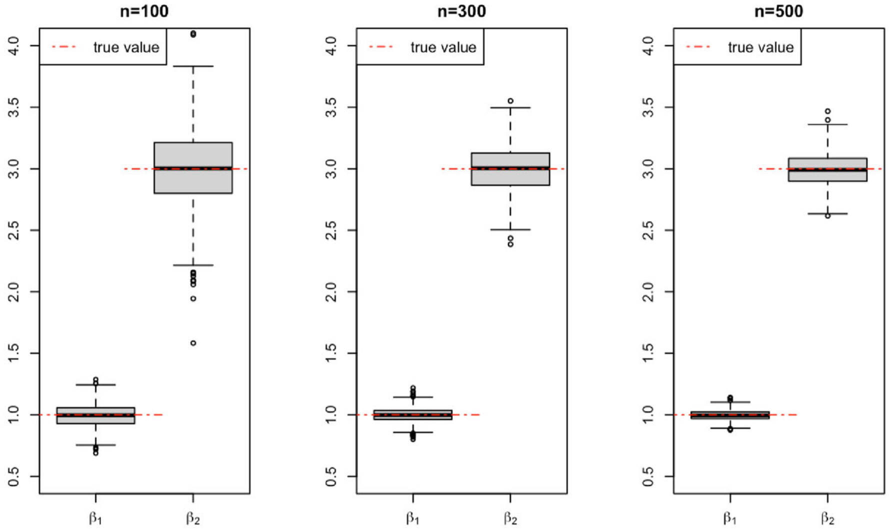

Fig. 2 Estimations of parameter $\beta = (\beta_1, \beta_2)^T$ with different sample sizes under the variance parameterization (i)

and sample sizes, with the bias in the estimates reducing for larger sample sizes. In Fig. 2, the medians of the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ in the three box plots are similar, but with the ranges of the boxes getting smaller when increasing n from 100 via 300 to 500. The effect is clearer visible for the $\hat{\beta}_2$'s than the $\hat{\beta}_1$'s since the larger magnitude of the true value of β_2 also comes with larger variability.

Similar simulations were conducted to test the accuracy under variance parameterization (ii), again using 1000 replicates of two-dimensional data from model (7) under each of three sample sizes of $n = 100$, $n = 300$, and $n = 500$. We report the

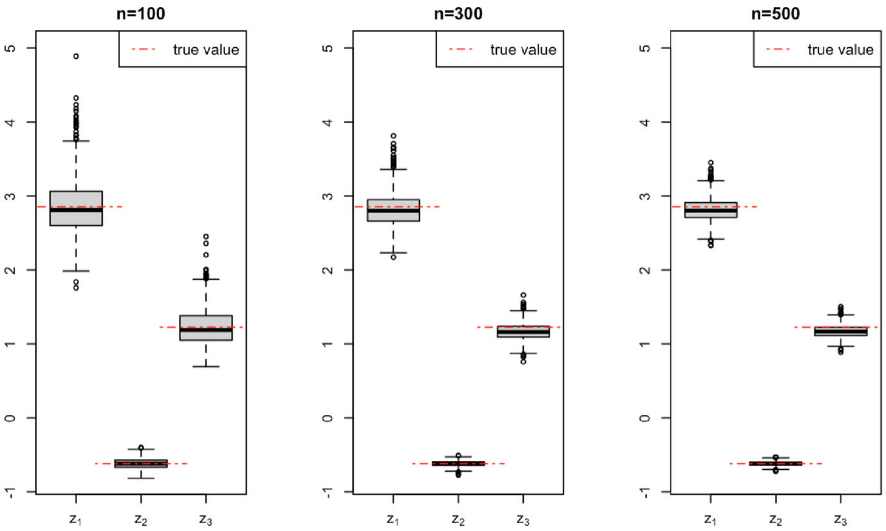

Fig. 3 Estimations of parameter z_k with different sample sizes under the variance parameterization (i)

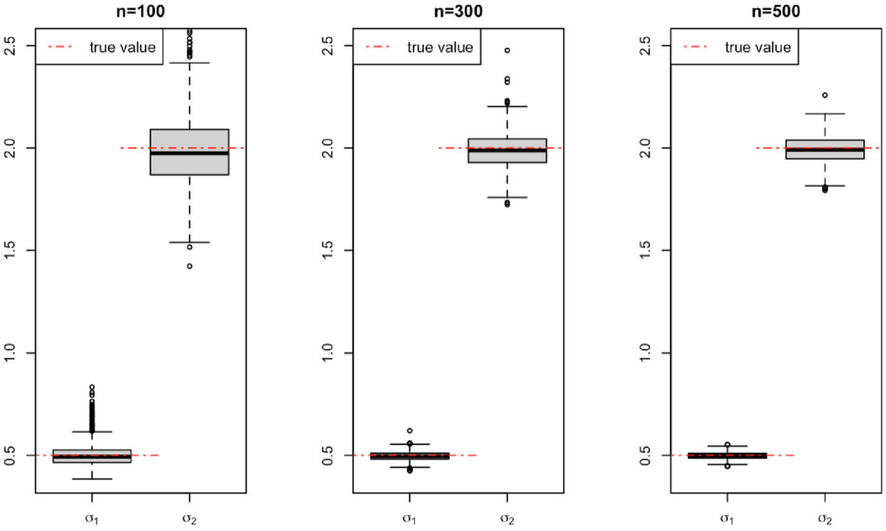

Fig. 4 Estimations of parameter σ with different sample sizes, where σ_1 and σ_2 are the diagonal components of the variance matrix, under the variance parameterization (i)

numerical results in Table 2 and display the estimated variance structures under this model in Fig. 5. We omit the boxplots for the other parameters as they are similar to those under parameterization (i). In 'Appendix B', we provide additional results and boxplots under parameterization (iii).

Overall, we can tell from the tables and figures that the estimators give sensible estimates of the parameters, the averaged estimates of the parameters are accurate

Table 2 Simulation results under variance parameterization (ii), where σ_{11} and σ_{21} are the diagonal elements of Σ_1, σ_{12} and σ_{22} are the diagonal elements of Σ_2

	True	Average estimates		
		$n = 100$	$n = 300$	$n = 500$
π_1	0.2000	0.2004	0.2001	0.2002
π_2	0.8000	0.7996	0.7999	0.7998
z_1	−0.5000	−0.5263	−0.5168	−0.4999
z_2	2.0000	2.0293	2.0182	2.0248
α_1	2.0000	2.0119	2.0024	2.0016
α_2	10.0000	10.0045	9.9998	9.9995
β_1	1.000	0.9929	0.9948	0.9955
β_2	3.000	2.9771	2.9871	2.9926
σ_{11}	0.2000	0.1972	0.1993	0.1998
σ_{21}	0.4000	0.3949	0.3971	0.3991
σ_{12}	1.0000	0.9614	0.9856	0.9948
σ_{22}	2.0000	1.9465	1.9880	1.9862

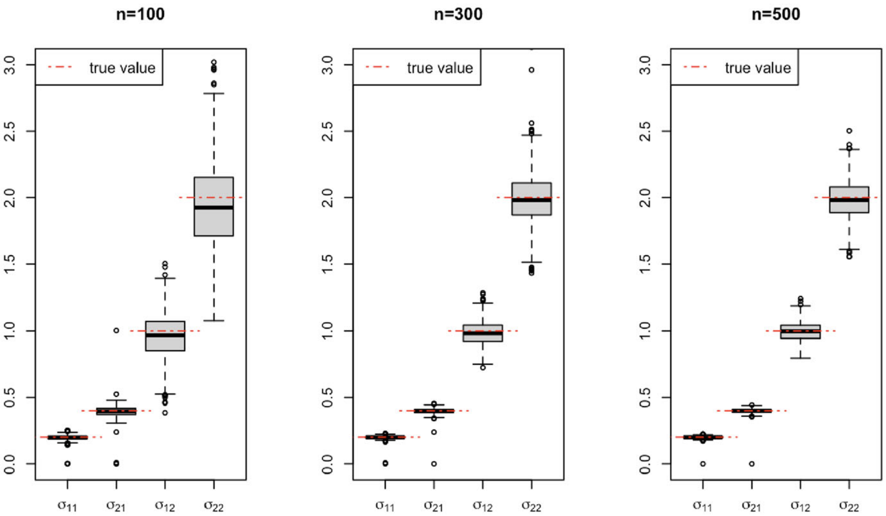

Fig. 5 Under variance parameterization (ii), estimations of variance parameters with different sample sizes, where σ_{11} and σ_{21} are the diagonal components of the variance matrix for mass point $k = 1$, σ_{12} and σ_{22} are the diagonal components of the variance matrix for mass point $k = 2$

compared to their true values, there appear to be no systematic biases, and the variability of the estimates reduces with increased sample size. The boxplots illustrate the consistency of estimators, where the boxes are squeezing to the true value (red horizontal line) as the sample size gets larger.

Next, we set up another set of simulations to address the importance of using the correct variance parameterization when fitting a model. For each model with each variance parameterization, we generate 200 replicates, each with sample size of 100, from the model. Then for the data generated from the model with variance parame-

Table 3 The number of estimated parameters used for AIC and BIC

Parameters	Variance (i)	Variance (ii)	Variance (iii)	Variance (iv)
π_k	$K-1$	$K-1$	$K-1$	$K-1$
z_k	K	K	K	K
α	m	m	m	m
β	m	m	m	m
Σ_k	m	mK	$\frac{m(m+1)}{2}$	$\frac{m(m+1)K}{2}$

terization (i), we fit the data to four different models, each with a different variance parameterization. For the remaining data sets generated from the model with variance parameterization (ii), (iii), and (iv), we do the same. We consider to use the AIC and BIC [18] as the model selection criteria, and we use the approximated likelihood (4) as the likelihood in AIC and BIC. For reference, the number of estimated parameters used in the calculation of AIC and BIC is shown in Table 3, where m is the dimension of data, and K is the number of mass points.

Figure 6 shows some key results: For the data sets generated from the model with variance parameterization (i), 73.5% of the fitted models with variance parameterization (i) lead to the smallest AIC values, and 95% of the fitted models with variance parameterization (i) lead to the smallest BIC values. For the data sets generated from the model with variance parameterization (ii), 88% of the fitted models with variance parameterization (ii) lead to the smallest AIC values, and 98% of the fitted models with variance parameterization (ii) lead to the smallest BIC values. For the data sets generated from the model with variance parameterization (iii), 87% of the minimum AIC values and 99% of the minimum BIC values are obtained from a fitted model with the variance parameterization (iii). For data sets generated from the model with variance parameterization (iv) and fitting the model with variance parameterization (iv), we obtain 99% of the minimum AIC values and 91.5% of the minimum BIC values. The results indicate that choosing a correct variance parameterization is significant for fitted model selection. Almohaimeed and Einbeck [6] discussed the use of AIC and BIC for model selection under NPML estimation. Although the BIC might lead to a different choice than AIC, Leroux [13] showed that using BIC for selecting the number of mixture components for finite mixture models is consistent. We use AIC and BIC as model selection criteria in our methodology.

4 Additional Inferential Aspects

In the previous sections, the focus was on estimating the parameters of model (7) from multivariate data $x_i \in \mathbb{R}^m$ and demonstrating that these estimators are (in an empirical sense) consistent, and the variance parameterizations are identifiable. In practice, these steps will rarely form an end in itself, but will be building blocks on the way to a more concrete statistical question. We now refer back to the four application pillars already mentioned in the introduction and explain these one by one. Additionally, we

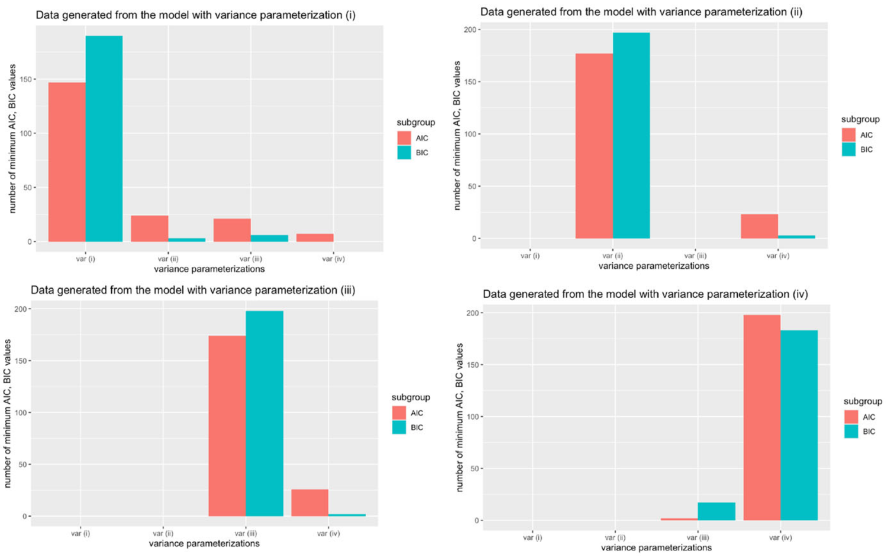

Fig. 6 Barplots showing the number of minimum AIC and BIC values obtained from fitted models with different variance parameterizations

will address the important question of how bootstrapped standard errors of covariate parameter estimates are obtained, and how these fare in comparison with univariate response models.

4.1 Clustering via MAP Estimation

We have already observed in Sect. 2.2 that the weights w_{ik} correspond to the posterior probability of observation i belonging to component k. The term 'posterior' is here be to be understood as the updated probability of class membership, having knowledge on the value of the observation x_i, as opposed to the 'prior' probability π_k, which does not make use of this information.

Given the availability of w_{ik} from the last iteration of the EM algorithm, observation x_i is then classified to the cluster $\hat{k}(x_i)$ to which it belongs with highest posterior probability,

$$\hat{k}(x_i) = \arg\max_k w_{ik}.$$

This cluster allocation rule is commonly known as maximum a posterior (MAP) rule. It is noted in this context that, after convergence of the EM algorithm, typically most w_{ik} are close to 0 or 1 (with obviously only one of them being close to 1), so that this allocation is in most cases very clear-cut. We will see examples for the application of the MAP rule in Sects. 5.1 and 5.3.

4.2 Dimension Reduction Through Predicted Latent Scores

One application of our methodology is the compression of m-dimensional data to one-dimensional, model-based scores, which can be considered as the summary information of the original data. This is achieved through the use of the 'projection'

$$z_i^* = \sum_{k=1}^{K} w_{ik} \hat{z}_k, \tag{24}$$

where $z_i^* \in \mathbb{R}$ [1]. Given the fitted model (1), z_i^* would be the best prediction of the position for the latent variable z_i that generates the original data x_i. Then, the following equation maps the one-dimensional scores back onto the higher dimensional original data space,

$$x_i^* = \alpha + \beta z_i^*,$$

where x_i^* is the compressed counterparts to the original data. It is clear that, unlike in, e.g. principal component analysis, the projections $x_i - x_i^*$ are not orthogonal to the linear subspace. However, they still can be meaningful: Under the given approach, all differences between observations to their cluster centres are treated as actual *errors*. The result of this is an increased robustness to such errors, as only clear deviations from a cluster lead to a projection beyond its centre. An example illustrating this behavior is provided in Fig. 9 in Sect. 5.1.

The one-dimensional scores, z_i^*, can then be used for subsequent inferential procedures, such as a predictor variable in a regression problem involving an external response variable y_i. This approach is illustrated by way of example in Sect. 5.2.

4.3 Ranking

The projected z_i^* provides a 'summary score' of all involved variables in the direction spanned by the latent line. Along this line, the positioning of the z_i^* is informative for the degree of which the variables jointly point into the direction of the latent variable. That is, high values of z_i^* would indicate overall high values of the contributing variables, and good agreement of what constitutes 'high'. For instance, if each of three variables constitutes price indexes for certain goods, then the higher these constituent indices are, the higher the overall price index will be. Hence, the order statstic of the z_i^*, denoted by $z_{[i]}^*$, can be used to rank the cases i, namely by $[i]$, $i = 1, \dots, n$. Many of these order statistics will be undistinguishable as the projections will be on (or close) to the same cluster centre. This makes sense from a clustering point of view: If observations cannot be distinguished statistically (i.e. if they are just distinguished by noise), their rank cannot be distinguished. De facto, in many cases, the $z_{[i]}$ will take as many *distinguishable* values as there are mass points. This concept will be explained in more detail by means of an example in Sect. 5.3.

4.4 Inclusion of Covariates

Where multivariate response data appear in statistical applications, the most common inferential approach is to define separate regression models for each of the individual variables constituting the multivariate response vector. For instance, while the linear model function `lm` in the statistical programming language R does allow for a multivariate response, the resulting fitted models correspond exactly to the individual one-dimensional response models. This approach, however, is ignoring the correlation of the different response variables, which, when taken into account, could lead to reduced parameter standard errors, and hence increased powers.

In the original model (1), $x_i \in \mathbb{R}^m$ can be explained by a one-dimensional coordinate system. Under the mixture representation of the model (7), certain latent groups along the one-dimensional line are driving the data generating process. However, these models do not yet allow for the presence of covariates in the data generating process of the x_i. To avoid confounding of the latent variable with such covariates (if they are known), the following is an extended model which includes a vector of p covariates related to the response variables,

$$x_i = \alpha + \beta z_i + \Gamma v_i + \varepsilon_i, \tag{25}$$

where $x_i \in \mathbb{R}^m$, $i = 1, 2, \dots, n$, $\alpha \in \mathbb{R}^m$, $\beta \in \mathbb{R}^m$, $v_i \in \mathbb{R}^p$ is the vector of the covariates, and $\Gamma_{m\times p}$ is a matrix of the coefficients of the covariates. The estimators of these parameters can be found in 'Appendix A'. When we have only one covariate in model (25), $v_i \in \mathbb{R}$, and we denote $\Gamma = \gamma \in \mathbb{R}^m$.

Notably, under model (25) with $x_i \in \mathbb{R}^m$, the 'models' for each of the m response variables would be linked through the random effect z_i, hence inducing correlation between units similar as for a multilevel model. An example for the use of this modelling technique is provided in Sect. 5.4.

4.5 Bootstrapped Standard Errors

In statistical practice, not only the estimation of Γ but also an assessment of its accuracy (or in other words, a quantification of its uncertainty) is of interest. Since the direct calculation of standard errors is generally difficult in the context of EM estimation, we propose a bootstrap procedure for their computation.

The bootstrap process is carried out with the following steps:

(i) We are given a data set $x_i \in \mathbb{R}^m$ and a covariate vector $v_i \in \mathbb{R}^p$, $i = 1, \dots, n$.
(ii) Fit the data x_i, v_i to model (25) to obtain the estimates of the parameters.
(iii) Sampling B data sets from model (25) with the estimated parameters obtained from (ii).
(iv) Fit these B data sets to our model and we would obtain B sets of $\hat{\gamma}$. Then, calculate the standard deviations across all B replicates of each of the $m \times p$ components of $\hat{\Gamma}$.

As an example, we generated a two-dimensional data set $x_i \in \mathbb{R}^2$ with $\pi = (0.3, 0.7)$, $z = (1.5, -0.6)$, $\alpha = (10, 2)$, $\beta = (1, 3)$, $\gamma = (0.5, 3)$ and $B = 300$.

Table 4 Estimations (Standard errors) of γ obtained using different methods

	$\hat{\gamma}_1$ (SE)	$\hat{\gamma}_2$ (SE)
`lm(.)`	0.5025 (0.2139)	2.8763 (0.4871)
Model (25) with bootstrap	0.4649 (0.1709)	2.7710 (0.3201)

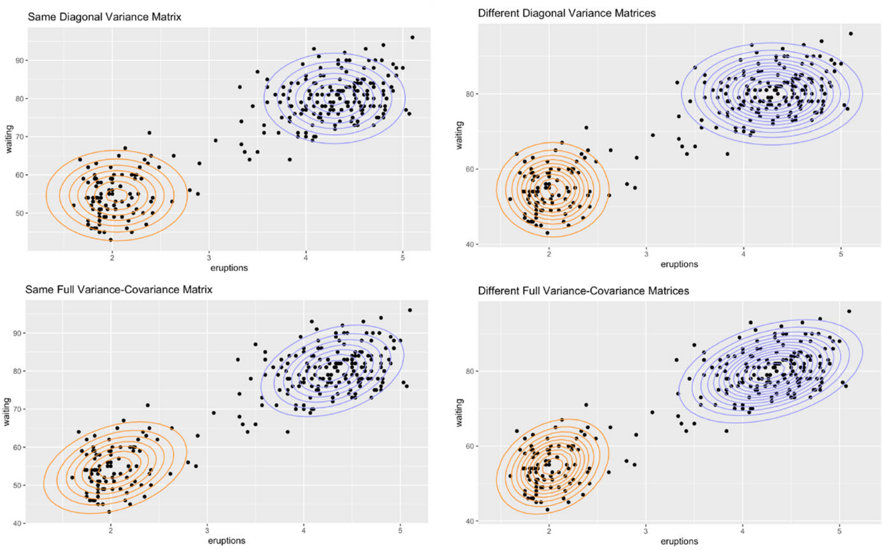

Fig. 7 Density contour plots with different variance parameterizations; top left (i); top right (ii); bottom left (iii); and bottom right (iv)

Then, we compared the estimates and standard errors obtained from the use of R function `lm` (when used as a multivariate response model), with those obtained from the procedures outlined in the previous and current subsection. The results are shown in Table 4; overall, our model leads to considerably smaller standard errors for the estimated coefficient parameter γ.

5 Applications

5.1 Faithful Data: Model Selection and Projection

In Sect. 2.2, we introduced four different variance parameterizations; here, we use again the faithful data set to illustrate the effect of using these different variance specifications on model fitting. Figure 7 shows the density contour plots for fitting the model with flexible variance parameterizations (i)–(iv). As shown from Table 5, the AIC and BIC values decrease when increasing the complexity of the variance parameterization, even though of course this does not need to be the case generally.

Table 5 AIC and BIC values for the faithful data under different variance parameterizations

	Variance (i)	Variance (ii)	Variance (iii)	Variance (iv)
AIC	2333.36	2317.61	2300.37	2286.53
BIC	2365.81	2357.28	2336.43	2333.40

The following are the parameter estimates from a fitted model with the selected parameterization (iv), i.e. different full variance–covariance matrices for each component: $\hat{\pi} = (0.3559, 0.6441)$, $\hat{\alpha} = (3.4878, 70.8971)$, $\hat{\beta} = (1.0788, 12.2038)$, $\hat{z} = (-1.3454, 0.7433)$, and

$$\hat{\Sigma}_1 = \begin{bmatrix} 0.0692 & 0.4352 \\ 0.4352 & 33.6973 \end{bmatrix}, \quad \hat{\Sigma}_2 = \begin{bmatrix} 0.1700 & 0.9406 \\ 0.9406 & 36.0462 \end{bmatrix}.$$

Figure 8 shows the clustering resulting from these estimates, according to the cluster allocation process that is described in Sect. 4.1.

We can obtain the scores (coordinates of the projected data along the one-dimensional subspace spanned by the latent variable) through the use of Eq. (24). We use the following images to illustrate the process of projecting the original data points onto the estimated low-dimensional space. In Fig. 1, the straight line is the one-dimensional latent space, and the red triangles positioned along the straight line are the estimated mixture centres $\hat{\alpha} + \hat{\beta}\hat{z}_k$. Figure 8 illustrates how the original data are assigned to different clusters following the MAP rule. The green points in Fig. 9 on the straight line are the compressed data, x_i^*, after projection onto that line. The most distinctive character between our methodology and the principal component analysis is that the projections are not orthogonal, which is shown in Fig. 10.

5.2 Soils Data: Dimension Reduction

In this example, we consider using the model based scores as the explanatory variable to fit a regression model with an additional new variable as the response variable. The data set we used for this analysis is the Soils data set in R package **carData** [11]. We construct a data frame with $n = 48$ and six variables: nitrogen, phosphorous, calcium, magnesium, potassium, and sodium (which are highly correlated, but do not all use the same units), and use an additional variable 'Density' (bulk density in gm/cm^3) as the response.

We apply the methodology laid out in Sect. 2.2 on the six-dimensional space of variables and use AIC and BIC to inform the choice of parameterizations and number of mass points. Details of the obtained AIC and BIC values using different number of mass points and variance parameterizations are shown in Tables 6 and 7. The AIC and BIC values given in these tables are the minimum values obtained over 20–50 runs with starting values chosen according to Sect. 2.5; the problem of finding the best solution gets harder when increasing k or the complexity of the error structure. We

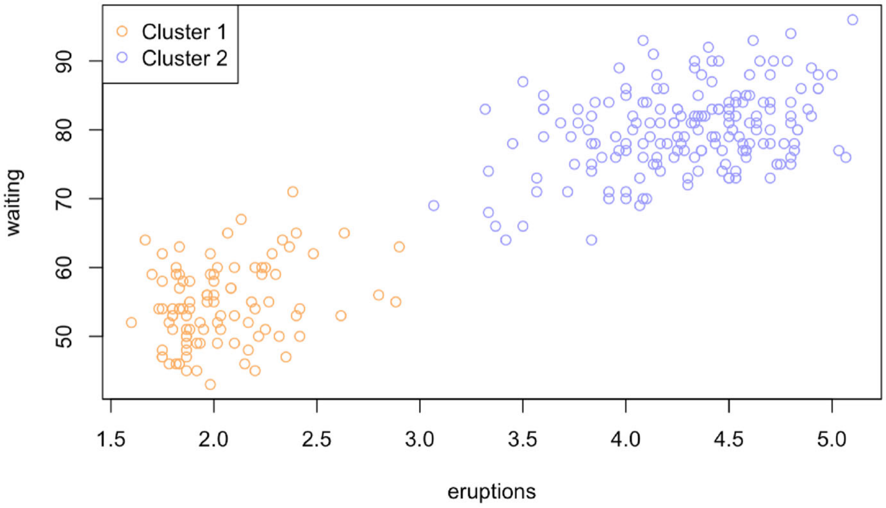

Fig. 8 For the faithful data, graph showing the original data points being assigned to different clusters according to the maximum a posterior (MAP) rule

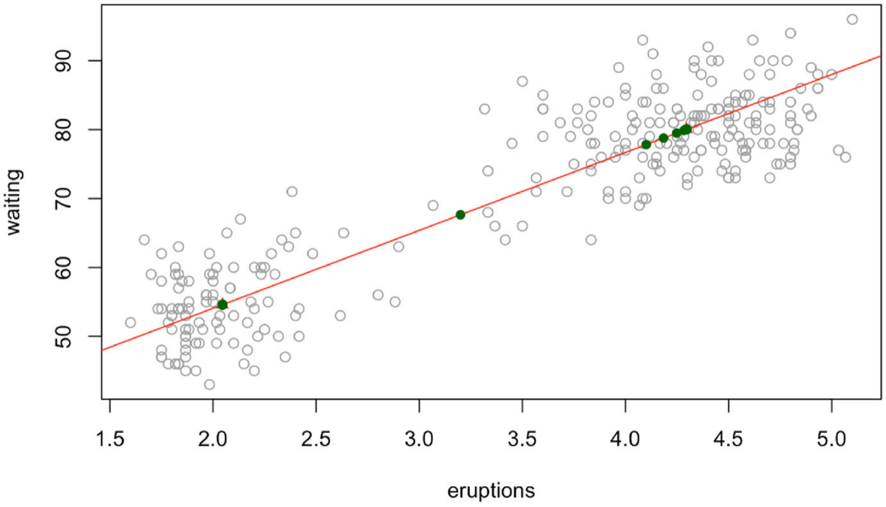

Fig. 9 For the faithful data, graph showing the projected data points x_i^* in green

find that AIC and BIC suggest to use variance parameterization (ii) with 4 mass points or 3 mass points, respectively, to fit the model.

Next we fit a regression model with the scores z_i^* being the predictor and the variable Density as response. Principal component regression is a commonly used technique for computing regressions when the explanatory variables are highly correlated. For a fair comparison, we construct the first principal component scores by projecting all data points onto the one-dimensional space and use these scores as the predictor. The fitted lines resulting from using two regression models are shown in Fig. 11. We see

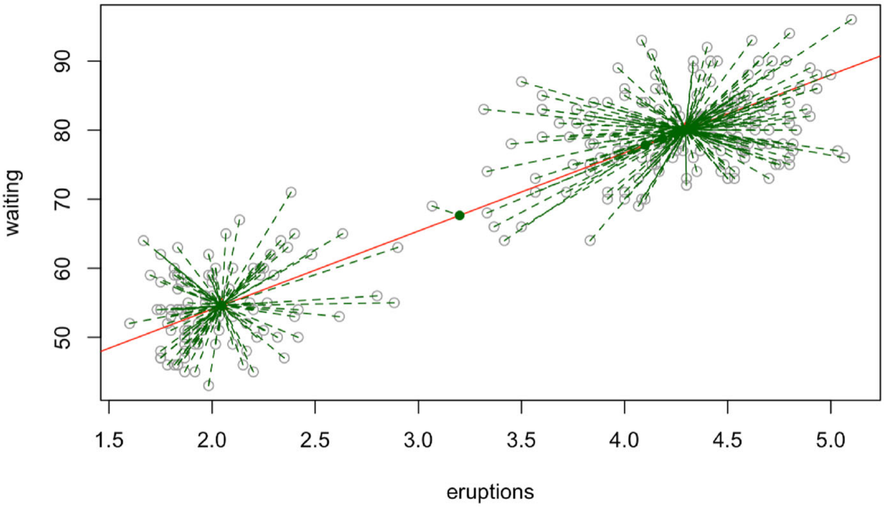

Fig. 10 For the faithful data, graph showing the projections of the original data points onto the estimated straight latent line

Table 6 AIC values for the Soils data under different variance parameterizations and different number of mass points

Variance parameterization	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
(i)	941.07	877.40	881.40	885.35	889.36
(ii)	888.38	827.99	**818.13**	823.82	849.45
(iii)	898.33	879.41	896.68	922.73	903.31
(iv)	842.40	940.30	876.08	826.69	NA

The smallest AIC value is in bold

Table 7 BIC values for the Soils data under different variance parameterizations and different number of mass points

Variance parameterization	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
(i)	980.37	934.84	928.18	935.87	943.62
(ii)	938.91	**893.49**	898.59	919.25	959.85
(iii)	965.70	950.51	971.53	1001.32	985.64
(iv)	949.06	1090.00	1068.81	1062.47	NA

The smallest BIC value is in bold

that the data are represented quite differently for our methodology. Table 8 shows the statistical measures that evaluate the performance of principal component regression in comparison with our approach (where we have considered both the AIC and BIC solutions). We find that our latent variable approach has a better performance for the non-scaled data. It is not unduly affected by scales or units and is robust concerning scaling.

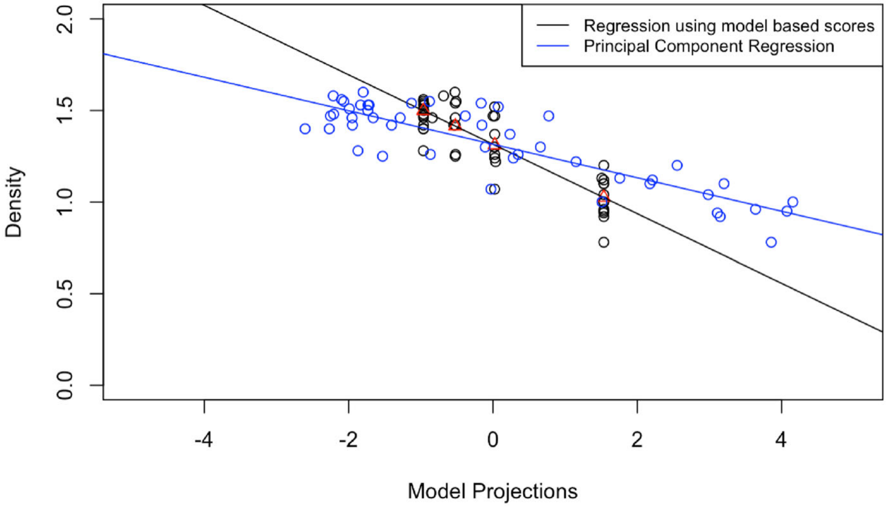

Fig. 11 Graph showing fitted lines using two regression models. In our methodology, the regression used model-based scores (model with $k = 4$) as the explanatory variable and another variable in the Soils data set called 'Density' as the response variable. The points in black correspond to Density values at the model-based scores, and points in red are the Density values predicted at the mass points. For the principal component regression, the fitted regression line in blue used the first principal component (the blue points) as the explanatory variable and the variable Density as the response variable

Table 8 Statistical measures of fit for the two regression models

Regression model	Non-scaled data	Scaled data
Latent variable model ($k = 3$)	R^2: 0.7430 RMSE: 0.1105	R^2: 0.7231 RMSE: 0.1137
Latent variable model ($k = 4$)	R^2: 0.7534 RMSE: 0.1084	R^2: 0.7457 RMSE: 0.1088
Principal Component Regression	R^2: 0.6226 RMSE: 0.1375	R^2: 0.7435 RMSE: 0.1097

5.3 Literacy Survey Data: Clustering and Ranking

League tables are produced for the comparison of different institutions. Aitkin et al. [3] compared student performance under different teaching techniques using variance component models. Aitkin and Longford [5] investigated several modelling approaches for the comparison of school effectiveness studies. Sofroniou et al. [19] used the International Adult Literacy Survey (IALS) data to construct league tables under the NPML estimation approach. In this section, we reconsider this data set for analysis. The International Adult Literacy Survey (IALS) was collected in 13 countries (or country-type entities) on Prose, Document, and Quantitative scales between 1994 and 1995. The data are reported as the percentage of individuals who could not reach a basic level of literacy (being the worst) in each country, the data can be found in 'Appendix C'.

As in [19], we only use the prose scale for the analysis. However, we take the separation of the reported prose results into male and female attainment differently into account than in that publication. We consider, for each of the 13 countries, male and female prose attainment as a bivariate response, allowing us to employ model (1) to describe the data, and so model (7) for parameter estimation. Since the gender variable is now being taken into account naturally in the response, no covariates at all are required in the model. Furthermore, since under this modelling approach, both female and male prose attainment for a given country are associated with the same random effect, it also eliminates the need to fit a two-level model as in [19] which is otherwise needed to correlate the female and male observations within each country. So, effectively, by using a gender-defined bivariate response, we are 'taking one level out' of the problem.

We fit the model with $k = 3$ mass points and with variance parametrization (ii) which leads to a minimum AIC value of 158.3963 and the smallest BIC value of 166.8705. The scores z_i^* are obtained as the posterior intercept and can be considered as the summary information of the original data. The task is here to rank the observations using the summary information. With the posterior probability matrix $W = (w_{ik})$ obtained at the convergence of the EM algorithm, upper-level units (countries) can then be classified into different clusters according to their largest posterior probabilities.

Table 9 shows the joint ranking of the countries, with the countries being classified into different clusters. In the table, the 3 mass points are ordered from left to right, from the cluster in which the country has the smallest percentage of adults being illiterate to the cluster in which the country has the largest percentage of adults being illiterate. The table shows that Sweden is assigned to mass point 1 which has the smallest number of people being illiterate. Poland is the only country that is assigned to the high illiteracy mass point 3. The Netherlands and Germany have posterior probabilities that spread across 2 mass points but are assigned to mass points 1 and 2 according to their highest posterior probabilities. We also fit the model (25) with $k = 5$ in order to compare to the results obtained by [19], the results and analysis can be found in 'Appendix C'.

5.4 Foetal Movement Data: Covariates and Standard Errors

We consider a set of foetal movements data collected before and during the COVID-19 pandemic. The study, which was executed by researchers of the Neonatal Research Lab at Durham University, aims to analyse the effects of COVID on foetal development [16]. The data were recorded via 4D ultrasound scans from a total of 40 mothers (20 before COVID and 20 during COVID) at 32 weeks gestation and consist of the number of movements each foetus carries out in relation to the recordable scan length. The ratio of these counts to scan length then forms the response variables of interest, with the following five specific movements recorded during the 4D ultrasound scans: upper face movements, head movements, mouth movements, touch movements, and eye blink. We are interested in the relationship of these five movements to the variable 'status', which indicates the period during which the data were collected ('pre-COVID' or

Table 9 Posterior intercepts and 'weight matrix' of posterior probabilities for the IALS data, with implied ranking ('league table'), for $k = 3$

Country	Posterior intercept	Mass points		
		0.154	0.769	0.077
		−1.325	−0.043	3.078
Sweden	−1.325	1.000		
Netherlands	−1.323	0.999	0.001	
Germany	−0.044	0.001	0.999	
Canada	−0.043		1.000	
Australia	−0.043		1.000	
Switzerland (French)	−0.043		1.000	
New Zealand	−0.043		1.000	
Belgium (Flanders)	−0.043		1.000	
Ireland	−0.043		1.000	
United States	−0.043		1.000	
Switzerland (German)	−0.043		1.000	
United Kingdom	−0.043		1.000	
Poland	3.078			1.000

Omitted entries correspond to 0.000

'during COVID'). For our analysis, this will be considered as a five-variate response, $x_i \in \mathbb{R}^5$ whereas status is the predictor, $v_i \in \mathbb{R}$.

We fit the data to model (25) with $k = 3$ and variance parametrization (ii) which leads to the smallest AIC (554.3622) value and BIC (613.473) value among all parametrizations and mass points. In principle, one could fit five separate linear regression models, each taking one of the movements score as the response and status as the predictor. We compare the estimates of the parameters and the parameter standard errors using this 'naïve' method to our proposed approach, using model (25), where the five equations are linked through a common random effect, the results are shown in Tables 10 and 11. Our methodology, involving a multivariate response model with random effect, gives parameter estimates which are consistent with the ones obtained from separate linear models, however enjoying reduced standard errors of the coefficients. The bottom row of Tables 10 and 11 also gives the p values of the estimated $\hat{\gamma}$'s. We observe that the p values also tend to be reduced, leading to a potentially different decision on the significance of a predictor variable if a decision threshold is crossed.

6 Conclusion

Multivariate data are rarely distributed homogeneously in space. In practice, one will often observe that the data reside on a latent linear subspace of a smaller dimension than itself, or that the data are concentrated into a certain number of clusters. From a statistical modelling point of view, these two concepts are usually dealt with in isolation or in succession, but not simultaneously. That is, often one will account for

Table 10 For the COVID data, estimations of γ obtained using individual linear models for upper face movements, head movements, mouth movements, touch movements, and eye blink

	Indiv. linear models				
	Upper face	Head movements	Mouth movements	Touch movements	Eye blink
Estimate ($\hat{\gamma}$)	0.472	0.217	2.600	0.317	0.367
Standard error	0.251	0.274	1.135	0.357	0.435
p value	0.068	0.432	0.028	0.380	0.405

Table 11 For the COVID data, estimations of γ obtained using the proposed multivariate response model with random effect

	Multivariate model				
	Upper face	Head movements	Mouth movements	Touch movements	Eye blink
Estimate ($\hat{\gamma}$)	0.460	0.203	2.549	0.297	0.346
Standard error	0.193	0.208	0.878	0.250	0.361
p value	0.051	0.381	0.048	0.224	0.323

Standard errors and *p* values are obtained via the bootstrap

the lower ‘intrinsic’ dimensionality through methods such as principal component analysis, partial least squares, factor analysis, etc., and then account for clustering in the resulting lower-dimensional space (for instance, by fitting a mixture model to the projections onto that space), or, less commonly, firstly partition the data into clusters and then apply separate compressions onto linear subspaces within each of them.

In this work, we have proposed a versatile statistical model based on a latent variable representation, which approaches both of these tasks simultaneously and enables solutions to a wide range of inferential problems, including multivariate regression problems in which the original data space might constitute either the predictors or responses. We have illustrated these scenarios, illuminating different inferential aspects, through a series of examples from various fields, hence illustrating the power of the proposed approach in statistical practice.

Our work has been based on the premise that the data set under investigation *does* feature latent structures which are worth of identifying or accounting for. The complexity of these latent structures is related to the choice of variance parameterization (i)–(iv). Empirical evidence for the identifiability of these variance matrices has been provided in the simulation section. From a practical point of view, we found variance parameterization (ii)—that is, diagonal, cluster-specific variance matrices–most useful, and also in fact selected by the AIC and BIC criteria in most cases. While the non-diagonal parameterizations (iii) and (iv) may be useful in certain situations, especially when the focus is on accurately describing the cluster structure as in Sect. 5.1, one could, at least conceptually, suspect that situations could arise where the latent variable and the variance matrices ‘compete’ for capturing the direction of the data cloud, hence potentially leading to non-identifiabilities in this respect. While we have not observed such problems in practical data sets, it is the case that convergence of the EM algorithm for scenarios (iii) and (iv) takes longe, and is also more sensitive to the selection of starting points.

As alluded to in the introduction, the basic concept behind the presented approach is not entirely new and has previously been expressed in the neural network community. However, those ideas have not been transferred into the statistical toolbox and embedded into a statistical modelling framework (as done here through the use of random effects) so far. It is also pointed out that several extensions of this work are possible, including the use of nonlinear or multivariate latent spaces with appropriate random effect specifications. Further, one could consider extending this framework towards non-Gaussian response distributions, requiring however more complex, GLM-type estimation methods.

We close with noting that our work can be considered as a particular type of multilevel (i.e. here, two-level) model, with the upper level corresponding to observations x_i and the lower level to 'measurements' x_{ij} on the 'repeated responses'. However, as we have seen in the example in Sect. 5.3, the shared random effect on the 'upper level' is directly obtained from the inferential framework without resorting to two-level ('variance component') modelling in a traditional sense. Spinning this thought further, the present methodology allows for a reduction of the number of levels in a genuine multilevel scenario. For instance, assume one has repeated measures of some quantity taken on the left and right ear of some individuals over time [21]. Then, rather than fitting a three-level model, the two ears could define the axes of a bivariate response model, reducing the problem to a two-level model. Work on such problems is in progress and will be reported elsewhere.

Acknowledgements We would like to thank Professor Nadja Reissland, Head of the Fetal and Neonatal Research Lab, Department of Psychology, Durham University, for making anonymised data available for re-analysis in our Sect. 5.4. We would also like to thank Dr Reza Drikvandi, Department of Mathematical Sciences, Durham University, for helpful comments on an earlier version of this work.

Funding No funding other than generic support by Durham University.

Data Availability Data used in Sects. 5.1 and 5.2 are part of publicly available R packages. Data from Sect. 5.3 are printed in the Appendix. Data from Sect. 5.4 will be made available by the authors in a forthcoming R package.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical Approval This publication does not present or analyse new data. Several existing data sets (which went through ethical approval where required when initially published) are analysed as secondary analyses. The data analyses as such do not require ethical approval.

A Estimators for Advanced Models

We recall model (25) introduced in Sect. 4.4:

$$x_i = \alpha + \beta z_i + \Gamma v_i + \varepsilon_i, \tag{25}$$

where $x_i \in \mathbb{R}^m$, $i = 1, 2, \ldots, n$, $v_i \in \mathbb{R}^p$, $\Gamma_{m \times p}$ is a matrix, $\varepsilon_i \sim N(0, \Sigma)$ is a Gaussian noise, and Σ in our methodology has four different parameterizations. The estimators used in the EM algorithm for this model would be obtained in the following:

With the probability density function for model (25),

$$f_{ik} = \frac{1}{(2\pi)^{m/2}} \frac{1}{|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x_i - \alpha - \beta z_k - \Gamma v_i)^T \Sigma_k^{-1} (x_i - \alpha - \beta z_k - \Gamma v_i) \right),$$

the log-likelihood would be:

$$\begin{aligned} l = & \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \log(\pi_k) + \sum_{i=1}^{n} \sum_{k=1}^{K} -\frac{1}{2} w_{ik} \log(|\Sigma_k|) + \sum_{i=1}^{n} \sum_{k=1}^{K} -\frac{m}{2} \log(2\pi) w_{ik} \\ & + \sum_{i=1}^{n} \sum_{k=1}^{K} -\frac{1}{2} w_{ik} (x_i - \alpha - \beta z_k - \Gamma v_i)^T \Sigma_k^{-1} (x_i - \alpha - \beta z_k - \Gamma v_i) \end{aligned}$$

E-step

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

M-step By taking partial derivatives of the log-likelihood with respect to each parameter, we obtain the score functions, and by equaling these score function to zero and solving them, we obtain the following estimators (already presented in the computationally efficient form as in Sect. 2.3):

$$\begin{aligned} \hat{\alpha} &= \frac{1}{n} \left(\sum_{i=1}^{n} x_i - \hat{\beta} \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \hat{z}_k - \sum_{i=1}^{n} \hat{\Gamma} v_i \right), \\ \hat{\beta} &= \frac{\begin{array}{c} \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \hat{z}_k x_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \hat{z}_k) - \sum_{i=1}^{n} \hat{\Gamma} v_i \sum_{k=1}^{K} w_{ik} \hat{z}_k \\ + \frac{1}{n}(\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \hat{z}_k)(\sum_{i=1}^{n} \hat{\Gamma} v_i) \end{array}}{\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \hat{z}_k^2 - \frac{1}{n}(\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \hat{z}_k)^2}, \\ \hat{z}_k &= \frac{\hat{\beta}^T \sum_{i=1}^{n} w_{ik} (x_i - \hat{\alpha} - \hat{\Gamma} v_i)}{\hat{\beta}^T \hat{\beta} \sum_{i=1}^{n} w_{ik}}, \end{aligned}$$

$$\hat{\Gamma} = \frac{\sum_{i=1}^{n} x_i v_i^T - \hat{\alpha} \sum_{i=1}^{n} v_i^T - \hat{\beta} \sum_{i=1}^{n} v_i^T \sum_{k=1}^{K} w_{ik} \hat{z}_k}{\sum_{i=1}^{n} v_i v_i^T}.$$

Since $\sum_{k=1}^{K} \pi_k = 1$, we need to apply a Lagrange multiplier by letting $\partial \left(l - \lambda(\sum_{k=1}^{K} \pi_k - 1)\right) / \partial \pi_k = 0$, then one obtains,

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

We let $\psi_i = (\psi_{i1}, \dots, \psi_{im})^T = \Gamma v_i \in \mathbb{R}^m$. Estimators for the flexible variance parameterizations are given as the following,

(i) $\Sigma = \text{diag}(\sigma_j^2)_{\{1 \le j \le m\}} \in \mathbb{R}^{m \times m}, k = 1, \dots, K,$

$$\hat{\sigma}_j^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} (x_{ij} - \hat{\alpha}_j - \hat{\beta}_j \hat{z}_k - \psi_{ij})^2$$

(ii) $\Sigma_k = \text{diag}(\sigma_{jk}^2)_{\{1 \le j \le m\}} \in \mathbb{R}^{m \times m}, k = 1, \dots, K,$

$$\hat{\sigma}_{jk}^2 = \frac{\sum_{i=1}^{n} w_{ik} (x_{ij} - \hat{\alpha}_j - \hat{\beta}_j \hat{z}_k - \psi_{ij})^2}{\sum_{i=1}^{n} w_{ik}}$$

(iii) $\Sigma = \Sigma_1 = \cdots = \Sigma_K \in \mathbb{R}^{m \times m}, k = 1, \dots, K,$

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} (x_i - \hat{\alpha} - \hat{\beta} \hat{z}_k - \hat{\Gamma} v_i)(x_i - \hat{\alpha} - \hat{\beta} \hat{z}_k - \hat{\Gamma} v_i)^T$$

(iv) $\Sigma_k \in \mathbb{R}^{m \times m}, k = 1, \dots, K,$

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{n} w_{ik} (x_i - \hat{\alpha} - \hat{\beta} \hat{z}_k - \hat{\Gamma} v_i)(x_i - \hat{\alpha} - \hat{\beta} \hat{z}_k - \hat{\Gamma} v_i)^T}{\sum_{i=1}^{n} w_{ik}}$$

B Additional Simulation Results

We provide here the figures arising from the simulation study carried out in Sect. 3, for variance parameterization (iii). We generate two-dimensional data from model (7) under three sample sizes of $n = 100$, $n = 300$, and $n = 500$, with 200 replicates.

The main results are summarized in Table 12 and Figs. 12, 13, and 14.

Table 12 Simulation results under variance parameterization (iii)

	True	Average estimates		
		$n = 100$	$n = 300$	$n = 500$
π_1	0.4000	0.4007	0.3983	0.3971
π_2	0.6000	0.5993	0.6017	0.6028
z_1	−0.8165	−0.8215	−0.8145	−0.8121
z_2	1.2247	1.2296	1.2320	1.2340
α_1	20.0000	20.0001	19.9965	19.9891
α_2	7.0000	7.0174	6.9780	6.9828
β_1	1.000	1.0065	0.9967	0.9980
β_2	3.000	2.9749	2.9864	2.9901

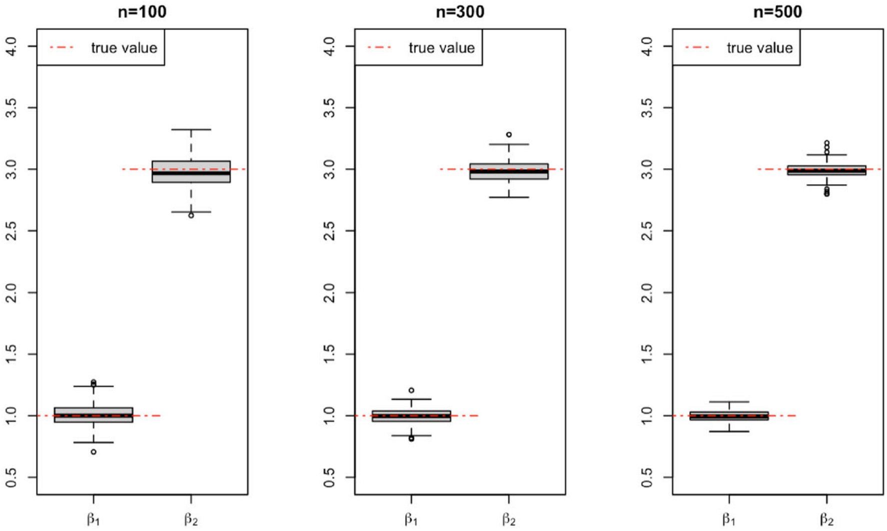

Fig. 12 Under variance parameterization (iii), estimations of parameter β with different sample sizes

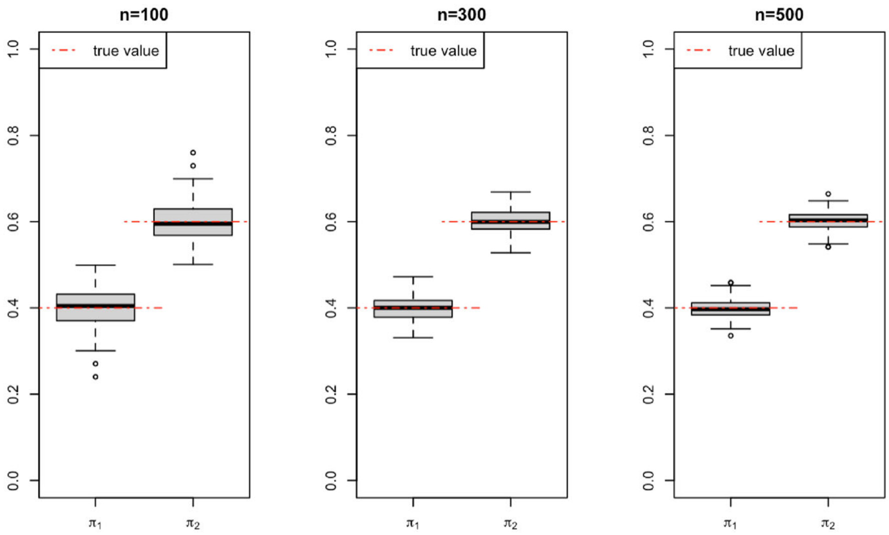

Fig. 13 Under variance parameterization (iii), estimations of parameter π_k with different sample sizes

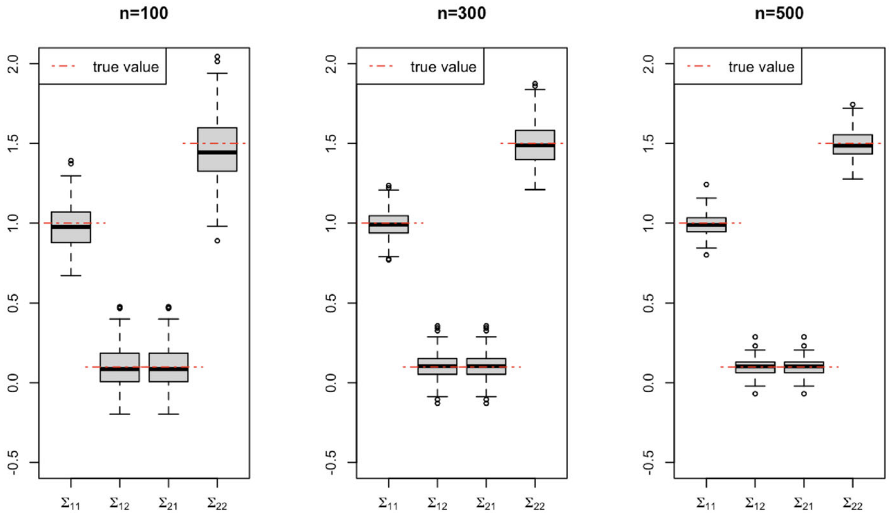

Fig. 14 Under variance parameterization (iii), estimations of parameter Σ with different sample sizes, where Σ_{11} and Σ_{22} are the diagonal elements, and Σ_{12} and Σ_{21} are the off-diagonal elements of the variance matrix. The true values are: $\Sigma_{11} = 1.0$, $\Sigma_{22} = 1.5$, $\Sigma_{12} = \Sigma_{21} = 0.1$

IALS Data

Table 13 displays the data used in Sect. 5.3, these data are the percentage of adults not reaching a basic level of prose in each country. Table 14 shows the ranking and clustering results obtained by [19], in which the data are modelled through a two-level binomial logit model with gender as a covariate using the `alldist` function from

Table 13 Proportion of adults not achieving prose level 2 in the IALS data set

Country	Male	Female
Ireland	24.21	20.93
United States	23.00	18.76
Switzerland (French)	17.46	19.44
Switzerland (German)	18.30	20.66
Canada	18.76	14.44
Belgium (Flanders)	15.55	21.61
Germany	14.31	13.31
United Kingdom	21.38	21.60
Netherlands	10.39	10.49
Poland	43.72	41.74
Sweden	7.31	7.18
Australia	18.33	15.69
New Zealand	19.94	16.52

Table 14 Classification and ranking for the IALS data in the paper by [19]

Country	Posterior intercept	Mass points				
		0.077 −2.602	0.093 −2.156	0.434 −1.599	0.319 −1.379	0.077 −0.330
Sweden	−2.60	1.00				
Netherlands	−2.16		1.00			
Germany	−1.72		0.21	0.79		
Australia	−1.60			1.00		
Canada	−1.59			0.97	0.03	
New Zealand	−1.58			0.92	0.08	
Belgium (Flanders)	−1.58			0.89	0.11	
Switzerland (French)	−1.54			0.72	0.28	
Switzerland (German)	−1.45			0.34	0.66	
United States	−1.38			0.01	0.99	
Ireland	−1.38				1.00	
United Kingdom	−1.38				1.00	
Poland	−0.33					1.00

the R package **npmlreg** [10]. For fair comparison, we fit the data (shown in Table 13) to model (25) with 5 mass points, i.e. $k = 5$, and the results are shown in Table 15.

Table 15 Posterior intercepts, weight matrix, and implied ranking for the IALS data using model (25) with $k = 5$

Country	Posterior intercept	Mass points				
		0.15	0.08	0.32	0.37	0.08
		−1.29	−0.66	−0.17	0.19	2.99
Sweden	−1.29	1.00				
Netherlands	−1.29	1.00				
Germany	−0.66		1.00			
Canada	−0.17			1.00		
Australia	−0.17			1.00		
New Zealand	−0.15			0.95	0.05	
Switzerland (French)	−0.03			0.61	0.39	
Switzerland (German)	0.04			0.42	0.58	
Belgium (Flanders)	0.15			0.12	0.88	
United Kingdom	0.19				1.00	
United States	0.19				1.00	
Ireland	0.19				1.00	
Poland	2.99					1.00

Omitted entries correspond to 0.00

References

1. Aitkin M (1996) Empirical Bayes shrinkage using posterior random effect means from nonparametric maximum likelihood estimation in general random effect models. In: Statistical modelling: proceedings of the 11th international workshop on statistical modelling, pp 87–94
2. Aitkin M (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. Stat Comput 6(3):251–262
3. Aitkin M, Anderson D, Hinde J (1981) Statistical modelling of data on teaching styles. J R Stat Soc Ser A (Gen) 144(4):419–448
4. Aitkin M, Francis B, Hinde J, Darnell R (2009) Statistical Modelling in R. Oxford University Press, Oxford
5. Aitkin M, Longford N (1986) Statistical modelling issues in school effectiveness studies. J R Stat Soc Ser A (Gen) 149(1):1–26
6. Almohaimeed A, Einbeck J (2022) Response transformations for random effect and variance component models. Stat Model 22(4):297–326
7. Bishop CM, Svensén M, Williams CK (1998) GTM: the generative topographic mapping. Neural Comput 10(1):215–234
8. Bouveyron C, Brunet-Saumard C (2014) Model-based clustering of high-dimensional data: a review. Comput Stat Data Anal 71:52–78
9. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B (Methodol) 39(1):1–22
10. Einbeck J, Darnell R, Hinde J (2018) npmlreg: nonparametric maximum likelihood estimation for random effect models. R package version 0.46-5
11. Fox J, Weisberg S, Price B (2020) carData: companion to applied regression data sets. R package version 3.0-4
12. Laird N (1978) Nonparametric maximum likelihood estimation of a mixing distribution. J Am Stat Assoc 73(364):805–811
13. Leroux BG (1992) Consistent estimation of a mixing distribution. Ann Stat 20:1350–1360
14. Murphy KP (2012) Machine learning: a probabilistic perspective. MIT Press, Cambridge

15. Panić B, Klemenc J, Nagode M (2020) Improved initialization of the EM algorithm for mixture model parameter estimation. Mathematics 8(3):373
16. Reissland N, Ustun B, Einbeck J (2023) The effects of lockdown during the covid pandemic on fetal movement profiles. Preprint on Research Square. https://europepmc.org/article/PPR/PPR735782
17. Sammel MD, Ryan LM, Legler JM (1997) Latent variable models for mixed discrete and continuous outcomes. J R Stat Soc Ser B (Stat Methodol) 59(3):667–678
18. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
19. Sofroniou N, Hoad D, Einbeck J (2008) League tables for literacy survey data based on random effect models. In: Proceedings of the 23rd international workshop on statistical modelling, Utrecht, pp 402–405. Statistical Modelling Society
20. Venables WN, Ripley B (2002) Modern applied statistics with S, 4th edn. Springer, New York. ISBN 0-387-95457-0
21. Verbeke G, Molenberghs G (2000) Linear mixed models for longitudinal data. Springer, New York