A multilevel multivariate response model for data with latent structures

Yingjuan Zhang[1], Jochen Einbeck[2], Reza Drikvandi[1]

[1] Department of Mathematical Sciences, Durham University, UK
[2] Durham Research Methods Centre, UK

E-mail for correspondence: `yingjuan.zhang@durham.ac.uk`

Abstract: We propose a two-level extension of a previously introduced multivariate latent variable model, which allows incorporating covariates on both levels. The presented model accounts for correlations among the response variables through univariate random effects which are modelled using a mixture distribution. We estimate the model parameters via an EM algorithm and provide simulation results and a real data application.

Keywords: Mixture distribution; Multivariate response model; Posterior intercepts; Random effects.

1 Introduction

The use of multivariate response models is not very widespread in statistical practice. This may be related to the circumstance that ready-to-use implementations are either only accessible via specialized software (such as SAS), or are equivalent to fitting separate univariate response models (such as R function `lm`). However, accounting for the multivariate response character has several inferential benefits including potentially increased powers. Zhang and Einbeck (2022) introduced a versatile latent variable model for dimension reduction and simultaneous clustering of multivariate data. However, their model did not allow for the inclusion of covariates and could not deal with repeated measures. This paper aims to provide such extensions. We consider a scenario where multivariate data $x_{ij} \in \mathbb{R}^m$ has 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

$$x_{ij} = \alpha + \beta z_i + \Gamma v_{ij} + \varepsilon_{ij}, \qquad (1)$$

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

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 coefficients, and ε_{ij} are independent Gaussian errors (if there is only one covariate, $v_{ij} \in \mathbb{R}$, we write $\Gamma = \gamma \in \mathbb{R}^m$). Under such a model, 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 z_i. Model (1) does not require the normality of random effects so no concerns to check the random-effects distribution (e.g., Drikvandi et al 2017).
Figure 1 illustrates a data scenario corresponding to this concept. The data used here is simulated from model (1) in the case that the latent variable obeys a three-point mixture distribution. The grey straight line represents the one-dimensional latent space $\alpha + \beta z$, and the black triangles positioned along the straight line the mixture centres of each component. The coloured thinner lines are for illustration only and show the trend of lower-level units within each each upper level (which is to some extent a result of the random error and to some part driven by the covariate). The orange triangles are the fitted values: $x^*_{ij} = \hat{\alpha} + \hat{\beta} z^*_i + \hat{\gamma} v_{ij}$, where $z^*_i = \sum_{k=1}^{K} w_{ik} \hat{z}_k \in \mathbb{R}$ are obtained as the posterior random effects using posterior probabilities of component membership w_{ik} (Aitkin, 1996).

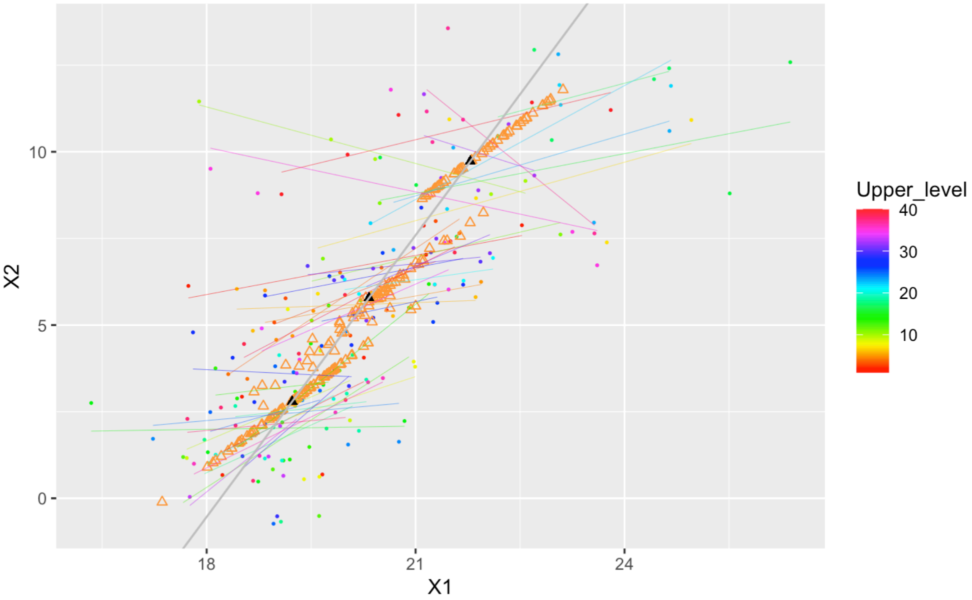

FIGURE 1. Simulated data with 40 upper level units, each with 5 lower level units, with $\alpha = (20, 10)$, $\beta = (1, 3)$, $\pi_k = (0.2, 0.3, 0.5)$, $z_k = (1.73, 0.29, -0.87)$, $\gamma = (0.5, 1)$. Observations are generated with component-specific diagonal variance matrices Σ_k. (We avoid the use of the term 'cluster' since this has a different connotation in the context of repeated measures.)

2 Methodology

We conduct the parameter estimation using maximum likelihood method. Since the component membership of each upper unit is unknown, we consider this as an 'incomplete data' problem, and apply the EM algorithm. The required complete data likelihood takes the shape

$$L_c = \prod_{i=1}^{r}\prod_{k=1}^{K}\left((\prod_{j=1}^{n_i} f_{ijk})\pi_k \right)^{G_{ik}},$$

where G_{ik} is an indicator variable taking the value 1 if upper unit i belongs to component k. We specify a multivariate Gaussian model for the component-specific densities f_{ik} in model (1) as

$$f_{ijk} = \frac{1}{(2\pi)^{m/2}}\frac{1}{|\Sigma_k|^{1/2}}\exp\left(-\frac{1}{2}(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij})^T\Sigma_k^{-1}(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij})\right).$$

The expected complete log-likelihood is then given by

$$l = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}\log(|\Sigma_k|) + \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi)w_{ik}$$
$$+\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij})^T\Sigma_k^{-1}(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij}),$$

where Σ_k is a component-specific diagonal variance matrix, and $w_{ik} = \frac{\pi_k f_{ik}}{\sum_l \pi_l f_{il}}$ is the probability of upper unit i belonging to component k. The computation of w_{ik} is via the E-step. The parameters α, β, z_k, Σ_k, and Γ will be estimated through the M-step. The key parameter estimates are:

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

and

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

3 Real data application

The real data used here is obtained from the International Adult Literacy Survey (IALS), collected in 13 countries 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 in each country. Based on the Prose scale only, Sofroniou et al (2008) used these data

TABLE 1. Posterior probabilities and intercepts for the IALS data. In the column 'mass points', the first two rows give estimated $\hat{\pi}_k$ and $\hat{z}_k$.

		Mass points			
		0.2308	0.5391	0.1532	0.0769
Country	posterior intercept	-1.1576	-0.0819	0.5904	2.8703
Sweden	-1.15760	1.0000	0.0000	0.0000	0.0000
Germany	-1.15756	1.0000	0.0000	0.0000	0.0000
Netherlands	-1.15754	0.9999	0.0001	0.0000	0.0000
Canada	-0.08188	0.0000	1.0000	0.0000	0.0000
Australia	-0.08188	0.0000	1.0000	0.0000	0.0000
Switzerland(French)	-0.08188	0.0000	1.0000	0.0000	0.0000
New Zealand	-0.08173	0.0000	0.9998	0.0002	0.0000
Belgium(Flanders)	-0.08163	0.0000	0.9996	0.0004	0.0000
Switzerland(German)	-0.08114	0.0000	0.9989	0.0011	0.0000
United States	-0.08036	0.0000	0.9977	0.0023	0.0000
Ireland	0.58386	0.0000	0.0098	0.9902	0.0000
United Kingdom	0.58912	0.0000	0.0019	0.9981	0.0000
Poland	2.87028	0.0000	0.0000	0.0000	1.0000

to rank countries according to their posterior intercepts $z_i^* = \sum_{k=1}^{K} \hat{z}_k w_{ik}$. We analyze the data considering the 3-variate response Prose, Document, and Quantitative, additionally including the lower-level covariate gender in the model; i.e. $m = 3$, $p = 1$ and $\Gamma = \gamma \in \mathbb{R}^3$.

The country-specific random effect z_i accounts for the correlation among the observations within upper-level units and the correlation among the three response dimensions of the model. We fit the model with $k = 4$ mass points and component-specific diagonal variances Σ_k, leading to an AIC value of 235.5 which does not drop significantly when increasing k further or with other variance parametrizations. Table 1 presents the joint ranking via the posterior random effect and classification of the countries. The table shows that Sweden, Germany, and the Netherlands are assigned to mass point 1 with the smallest number of people being illiterate. Poland is the only country that is assigned to the high illiteracy mass point 4. The US and Ireland have posterior probabilities that spread across two mass points but are assigned to different components. Using all three measurements as a multivariate response, the component allocation of each country is more decisive compared to the results using just Prose (Sofroniou et al, 2008).

4 Simulation study

We conduct a simulation study to examine the performance of our method. 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 in the parameter estimates. 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 generate 200 replicates from the model (1) with one lower level covariate in all three scenarios, with the covariate generated from a normal distribution with a mean of 0.3 and a standard deviation of 0.2. The results indicate that when we increase the upper-level units, the parameters' RMSE decreases stronger than when increasing the lower-level units. Then we further increase the upper level units to be $r = 200$ and keep the lower level units $n_i = 5$ for $i = 1, 2, \ldots, r$. The key results are shown in Figure 2, Table 2 and Table 3.

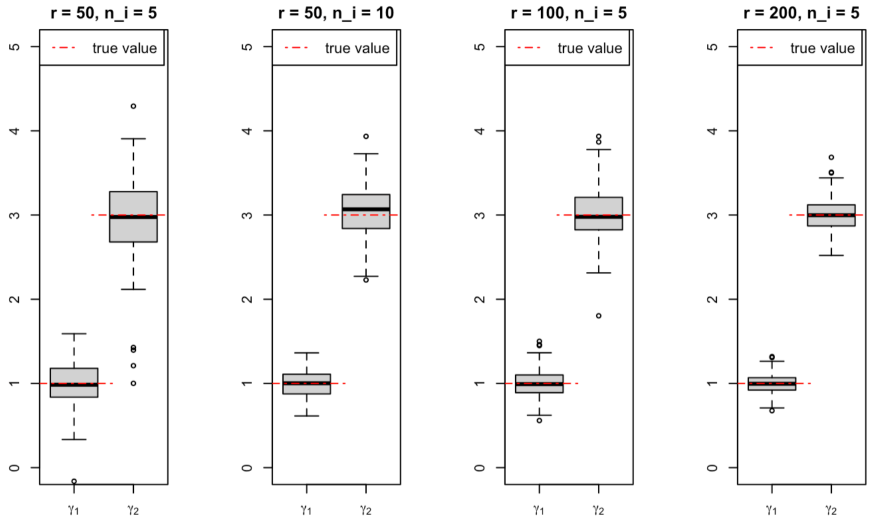

FIGURE 2. Estimates of key parameter γ with different number of upper-level and lower-level units.

5 Conclusion

This paper provides an extended random effect model that applies to two-level multivariate response data with latent structures. An EM algorithm is used for parameter estimation. In particular, a nonparametric maximum likelihood method (Aitkin 1999) is used for estimation of the random effect where the mass points z_k and their weights π_k, $k = 1, 2, \ldots, K$ are treated as unknown parameters to be estimated in the EM algorithm. An application of constructing a league table using the IALS data is provided. Another application is to fit multivariate response models. **Note:** A typo

TABLE 2. Estimates of key parameters γ, z_k and α with different upper-level and lower-level units.

		Average estimates			
	True	$r = 50, n_i = 5$	$r = 50, n_i = 10$	$r = 100, n_i = 5$	$r = 200, n_i = 5$
γ_1	1.000	0.989	0.993	0.991	0.995
γ_2	3.000	3.036	2.972	3.009	2.998
z_1	-0.816	-0.807	-0.814	-0.820	-0.809
z_2	1.225	1.268	1.258	1.234	1.246
α_1	2.000	2.037	2.039	2.034	1.991
α_2	10.000	10.020	10.007	10.019	10.002

TABLE 3. RMSE for key parameters γ, z_k and α with different 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	0.269	0.167	0.166	0.118
γ_2	0.474	0.297	0.296	0.194
z_1	0.124	0.124	0.084	0.068
z_2	0.198	0.207	0.133	0.132
α_1	0.464	0.447	0.302	0.232
α_2	0.172	0.161	0.120	0.088

in the published proceedings version in the expression for L_c has been fixed in this version.

References

Aitkin, M. (1996). Empirical bayes shrinkage using posterior random effect means from nonparametric maximum likelihood estimation in general random effect models. In: *Proc's of the 11th IWSM*, Orvieto, Italy, 87 – 94.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55(1), 117 – 128.

Drikvandi, R., Verbeke, G. and Molenberghs, G. (2017). Diagnosing misspecification of the random-effects distribution in mixed models. *Biometrics*, 73(1), 63 – 71.

Sofroniou, N., Hoad, D., and Einbeck, J. (2008). League tables for literacy survey data based on random effect models. In: *Proc's of the 23rd IWSM*, Utrecht, Netherlands, 402 – 405.

Zhang, Y., and Einbeck, J. (2022). Simultaneous linear dimension reduction and clustering with flexible variance matrices. In: *Proc's of the 36th IWSM*, Trieste, Italy, 612 – 617.