Simultaneous linear dimension reduction and clustering with flexible variance matrices

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

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

[1,2] Durham Research Methods Centre, UK.

1. Introduction

Random effect methodology is proposed for the dimension reduction of multivariate data, $x_i \in \mathbb{R}^m$. This is achieved by projecting the original data onto the estimated low-dimensional latent space, $\alpha + \beta z$, where $\alpha, \beta \in \mathbb{R}^m$ and z is a one-dimensional random effect represented by a discrete mixture with mass points $z_1,...,z_k$ and masses $\pi_1,...,\pi_k$, $k = 1, ..., K$. The observed data are assumed to be generated from the 'generative linear mixture model' (Lawson and Einbeck, 2012)

$$x_i = \alpha + \beta z_k + \varepsilon_i,$$

where $\alpha + \beta z_k$ are the cluster centers on the straight line, and $\varepsilon_i \sim N(0, \Sigma)$ is the Gaussian noise added to the cluster centers. Under the original approach, the variance matrix $\Sigma \in \mathbb{R}^{m\times m}$ is assumed to be a diagonal matrix, $\text{diag}(\sigma_j^2)_{\{1\le j\le m\}}$ and to be the same for all K components of the mixture. The previous assumption on the variance disregards other geometric features that clusters might have, such as clusters with different sizes, shapes or orientations determined by the covariance. So, we consider several types of variance matrix parametrizations. We also solve an identifiability issue inherent to the original approach.

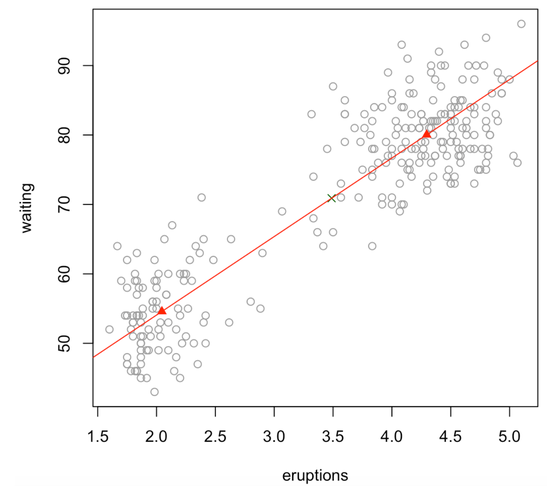

Fig 1. **Faithful Data**

2. Methodology

- The parameters $\alpha, \beta, z_k, \sigma_j$ and π_k will be estimated through the EM Algorithm. By using the posterior probability that x_i belongs to component k,

$$w_{ik} = \frac{\pi_k f_{ik}}{\sum_{l=1}^{K} \pi_l f_{il}}$$

where for the (original) generative linear mixture model

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

one obtains the corresponding (expected) complete log-likelihood,

$$l = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \log \pi_k + w_{ik} \log f_{ik}.$$

- The following are the estimators when using the variance parametrizations, with (ii) to (iv) being new contributions of this work,

(i) $\Sigma \in \mathbb{R}^{m\times m}$, $\text{diag}(\sigma_j^2)_{\{1\le j\le m\}}$, $\quad \hat\sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_{ij} - \hat\alpha_j - \hat\beta_j \hat z_k)^2$

(ii) $\Sigma_k \in \mathbb{R}^{m\times m}$, $\text{diag}(\sigma_{jk}^2)_{\{1\le j\le m\}}, k = 1, ..., K$, $\quad \hat\sigma_{jk}^2 = \frac{\sum_{i=1}^{n} w_{ik}(x_{ij} - \hat\alpha_j - \hat\beta_j \hat z_k)^2}{\sum_{i=1}^{n} w_{ik}}$

(iii) $\Sigma \in \mathbb{R}^{m\times m}$, $\Sigma = \Sigma_1 =, ..., = \Sigma_k$, $\quad \hat\Sigma = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_i - \hat\alpha - \hat\beta \hat z_k)(x_i - \hat\alpha - \hat\beta \hat z_k)^T$

(iv) $\Sigma_k \in \mathbb{R}^{m\times m}$, $k = 1, ..., K$, $\quad \hat\Sigma_k = \frac{\sum_{i=1}^{n} w_{ik}(x_i - \hat\alpha - \hat\beta \hat z_k)(x_i - \hat\alpha - \hat\beta \hat z_k)^T}{\sum_{i=1}^{n} w_{ik}}$

3. Identifiability

There is a product term of βz_k in the original model, which makes the parameters β, z_k unidentifiable. Furthermore, also α is unidentifiable as the same model could be attained by translating all z_k's along the line. In order to fix this identifiability problem, we standardize z_k, by letting

$$\sum_{k=1}^{K} \pi_k z_k = 0, \qquad \sum_{k=1}^{K} \pi_k z_k^2 - (\pi_k z_k)^2 = 1,$$

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). Additionally, to identify the direction of the latent variable, we enforce $\hat\beta_1 \ge 0$.

4. Simulation

A simulation is set up to test the correctness of the methodology, after implementing the identifiability fixes, under variance parametrization (i). We use 2-dimensional data with three individual sample sizes $n = 100$, $n = 300$, and $n = 500$, and generate 1000 data sets for each sample size. Then we compare the average estimated values to the true values of the parameters used to generate these data, the results are shown in table below. Most biases are around 0.005, and no biases greater than 0.05. The estimated parameters are getting closer to the true values as the sample size gets larger.

Table 1: Simulation Results for β, α and z.

	β_{true}	α_{true}	z_{true}
	(1.0000, 3.0000)	(-1.0000, 1.0000)	(2.8023, 1.1675, -0.6171)
n	$\bar\beta_{est.}$	$\bar\alpha_{est.}$	$\bar z_{est.}$
100	(0.9915, 2.9974)	(-0.9936, 1.0235)	(2.8547, 1.2262, -0.6186)
300	(0.9986, 2.9982)	(-0.9985, 1.0036)	(2.8130, 1.1693, -0.6193)
500	(0.9966, 2.9899)	(-0.9985, 0.9983)	(2.8119, 1.1708, -0.6191)

5. Application

- From the Soils data set in R package **car**, we construct a data frame of six variables: Nitrogen, Phosphorous (in ppm), Calcium, Magnesium, Potassium (in me/100 gm) and Sodium. The features in this data frame are on wildly different scales and in different units. We apply the methodology with variance parametrization (ii). Fitting a model with k=4 mass points leads to an AIC value of 823.34. We obtain projected data points by $x_i' = \sum_{k=1}^{K} w_{ik}\hat z_k$ (Aitkin, 1996). We fit a linear regression model with the variable Density (in gm/cm^3) as the response variable and the projected data as the predictor. For fair comparison, we construct the first principal component scores by projecting all data points onto the 1-dimensional space and use these scores as the predictor.
- Table 2 shows the statistical measures evaluating the performance of the two models. We find that our approach performs better for the non-scaled data, and that both approaches perform similarly for the scaled data.

Regression Model	Non-scaled Data	Scaled Data
Mixture-based Approach	R^2 : 0.7534	R^2 : 0.7457
	RMSE : 0.1084	*RMSE* : 0.1096
Principal Component Regression	R^2 : 0.6226	R^2 : 0.7435
	RMSE : 0.1378	*RMSE* : 0.1099

Table 2. **Statistical measures of fit for the two regression models.**

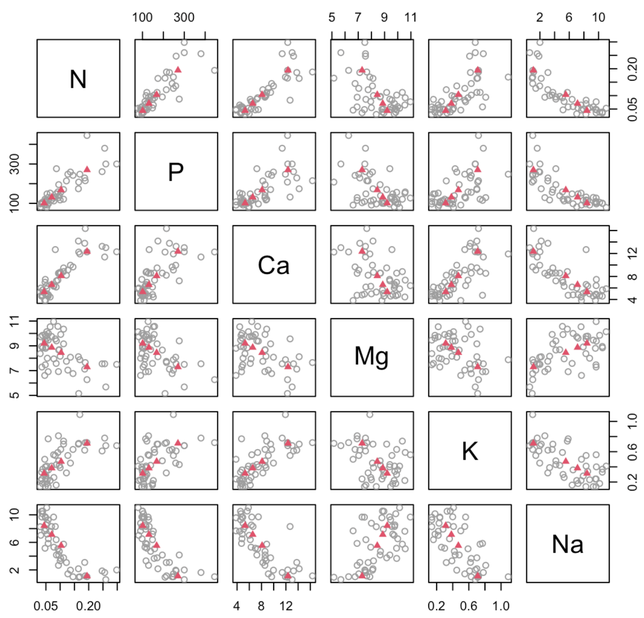

Fig 2. **Soils Data, cluster centers**

6. 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*, pp 87–94, Orvieto, Italy.

Lawson, A. and Einbeck, J. (2012)Generative linear mixture modelling. In: *Proceedings of the 27th International Workshop on Statistical Modelling*, pp 595–600, Prague, Czech Republic.

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