Generative Linear Mixture Modelling

Antony Lawson[1], Jochen Einbeck[1]

[1] Department of Mathematical Sciences, Durham University, Durham, DH1 3LE, England

E-mail for correspondence: jochen.einbeck@durham.ac.uk

Abstract: For multivariate data with a low–dimensional latent structure, a novel approach to linear dimension reduction based on Gaussian mixture models is proposed. A generative model is assumed for the data, where the mixture centres (or 'mass points') are positioned along lines or planes spanned through the data cloud. All involved parameters are estimated simultaneously through the EM algorithm, requiring an additional iteration within each M-step. Data points can be projected onto the low–dimensional space by taking the posterior mean over the estimated mass points. The compressed data can then be used for further processing, for instance as a low–dimensional predictor in a multivariate regression problem.

Keywords: EM; Dimension Reduction; Mixture Modelling.

1 Introduction

Mixtures of exponential family distributions are often used to model complex data structures, with finite Gaussian mixtures being the most common representant of such models. In this article we are interested in situations where a multivariate data set, $x_i \in \mathbb{R}^m$, $i = 1, \ldots, n$, possesses a latent structure of lower dimension $d < m$ (these 'data' may play the role of a 'predictor space' in a multivariate regression problem, but this is not relevant for the moment). The objective, for now, is to recover the latent structure, and to compress the original data by projecting them (in some form) onto the estimated latent space. As a first step towards a more general handling of this problem, we consider a simplified scenario in which the latent structure is thought to be a straight line, say $\alpha + \beta z$, with $\alpha, \beta \in \mathbb{R}^m$, $z \in \mathbb{R}$, through an m-dimensional space. The variable z is considered as a random effect, and represented by a discrete distribution with mass points $z_k \in \mathbb{R}$ and masses $\pi_k, k = 1, \ldots, K$. The data are assumed to be generated by adding Gaussian noise $\varepsilon_i \sim N(0, \Sigma)$ to mixture centres $\alpha + \beta z_k \in \mathbb{R}^m$ positioned along this line, yielding the generative linear mixture model

$$x_i = \alpha + \beta z_k + \varepsilon_i. \tag{1}$$

The variance matrix $\Sigma \in \mathbb{R}^{m\times m}$ is assumed to be of diagonal form $\mathrm{diag}(\sigma_j^2)_{\{1\le j\le m\}}$, and to be the same for all K components of the mixture.

2 The EM Algorithm

As for univariate mixtures, the data likelihood, L, can be written in the form

$$L = \textstyle\prod_{i=1}^{n}\sum_{k=1}^{K} f_{ik}\pi_k$$

where, for model (1),

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

In order to setup an EM algorithm, we need to consider the complete data likelihood, which is the likelihood of the data given that we know the component each x_i belongs to. However, the components each datum belongs to are unobservable, so we must use the posterior probabilities that x_i belongs to component k, which are obtained as

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

The complete data likelihood therefore takes the form

$$L^* = \textstyle\prod_{i=1}^{n}\prod_{k=1}^{K}(f_{ik}\pi_k)^{\omega_{ik}},$$

giving the complete log-likelihood

$$\ell^* = \log(L^*) = \sum_{i=1}^{n}\sum_{k=1}^{K}\omega_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K}-\frac{1}{2}\omega_{ik}\log(|\Sigma|)$$

$$+\sum_{i=1}^{n}\sum_{k=1}^{K}-\omega_{ik}\frac{m}{2}\log(2\pi)+\sum_{i=1}^{n}\sum_{k=1}^{K}-\frac{1}{2}\omega_{ik}(x_i-\alpha-\beta z_k)^T\Sigma^{-1}(x_i-\alpha-\beta z_k)$$

Score equations were obtained by partially differentiating ℓ^* with respect to each of the variables. Although an analytical solution was not obtained for α, β and z_k, we were able to find an iteration process involving these parameters. Solving the score equations for α and z_k give

$$\hat{z}_k = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\sum_{i=1}^{n}\omega_{ik}x_{ij}}{\sum_{i=1}^{n}\omega_{ik}} - \hat{\alpha_j}\right)/\hat{\beta}_j$$

(with the subscript j denoting the $j-$th component of the respective vector), and

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

Substituting $\hat{\alpha}$ into the equation for $\hat{\beta}$ and solving gives

$$\hat{\beta} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K}\omega_{ik}x_i\hat{z}_k - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}\sum_{k=1}^{K}\omega_{ik}\hat{z}_k\right)}{\sum_{i=1}^{n}\sum_{k=1}^{K}\omega_{ik}\hat{z}_k^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\sum_{k=1}^{K}\omega_{ik}\hat{z}_k\right)^2}.$$

To implement this in the EM algorithm, at each M-step there will be an internal iteration loop involving these parameters. First, the $\hat{z}_k$ will be calculated using the values of the previous internal iteration. Then $\hat{\beta}$ will be calculated using the newly calculated values of $\hat{z}_k$. Then finally $\hat{\alpha}$ will be calculated using the new values of $\hat{\beta}$ and $\hat{z}_k$. The initial $\hat{\beta}$ and $\hat{\alpha}$ values used will be those from the previous E-step.
Given the new values of $\hat{\alpha}, \hat{\beta}$ and $\hat{z}_k$, the score equation for $\hat{\sigma}_j$ solves very easily to

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

Using a Lagrange multiplier, $\ell^* - \lambda(\sum_{k=1}^{K}\pi_k - 1)$, one obtains

$$\hat{\pi}_k = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\omega_{ik}.$$

3 Results

Analysis was carried out on the mussels data set (Bura and Cook, 2001; available from R package **dr**), considering intially the data frame constituted by the variables shell length, shell width, shell height, and shell mass. Applying the above methodology, Table 1 shows how the disparity, $-2\log L$, of the model varies with number of components, K. The disparity decreases considerably with each component added, until the 8th component, where the disparity stabilizes at a value of 2608.088.
A bootstrapping method was required to test for a sensible number of components. Testing a model with 5 components against one with 6 returned a p-value of 0.31, and testing 4 components against 5 returned a p-value of 0.01, implying a 5 component model is a good representation of the data.
The iteration loop in the M-step converges very fast, with not more than 5 iterations initially, quickly falling to 3 iterations after a few EM cycles.
The number of EM iterations taken for the variables to convergence was also observed and the $\hat{\sigma}_j$ were generally the fastest to converge, with $\hat{\beta}$ converging slower, $\hat{\alpha}$ and $\hat{z}_k$ a little slower than $\hat{\beta}$, and $\hat{\pi}_k$ considerably

TABLE 1. Table of measurements for a variety of components

K	Disparity	RSS	R^2	# Iterations for disparity to converge
2	2881.936	7.057	0.6421	7
3	2804.671	5.574	0.7767	26
4	2676.017	5.222	0.8041	13
5	2645.342	5.073	0.8151	34
6	2630.438	5.010	0.8196	102
7	2623.526	4.783	0.8356	126
8	2608.088	4.759	0.8373	145
9	2608.088	4.759	0.8373	194

slower. The disparity of the models converge somewhat faster than any of the components.

The next step taken in the analysis was projecting the data points onto the fitted line. For each data point x_i, projected (or compressed) data are obtained as 'posterior means' (Aitkin, 1996)

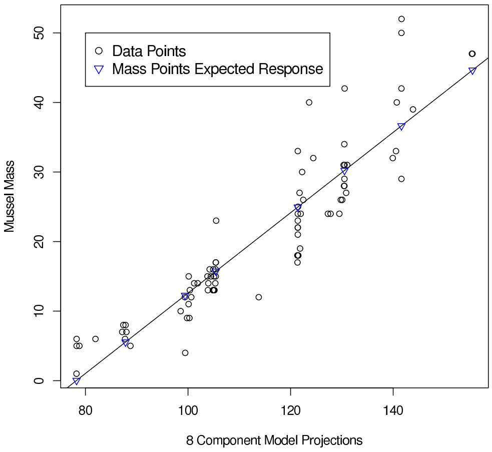
$$x_i^P = \sum_{k=1}^{K} \omega_{ik} \hat{z}_k.$$

These 'projections' are not orthogonal, and hence are of fundamentally different character as those in PCA, for instance. To verify the usefulness of this type of compression, we considered now the additional variable mussel mass as response variable, y, and fitted a simple linear regression model for y_i versus x_i^P. The resultant line is shown along with $(x_i^P, y_i), i = 1, \ldots, n$ in Figure 1 and appears to represent the data reasonably well. The RSS and R^2 values for the fitted linear model were recorded for each model and are included in Table 1. It is clear that the model improves as number of components is increased. Comparing these results to the 'parametric inverse regression' method by Bura and Cook (2001), with RSS = 6.051 and $R^2 = 0.741$, we find the proposed mixture–based approach to perform considerably better.

4 Discussion

This article has reported on the results of a pilot study using the most simple of all latent model scenarios, namely a straight line spanned through the data which carries the mixture centres. This research has been tentatively extended in two directions: Firstly, the case of a bivariate latent structure (i.e., a plane), and secondly, the case of a 'staggered line' which is allowed to change its direction at each mass point. In both cases, the likelihood

FIGURE 1. Graph of mussel mass (response) modelled by projection index (x_i^P).



equations were still tractable and the algorithms converged in reasonable time, though the issue of starting point selection for the EM algorithm requires more attention with increasing complexity of the model.
The presented work could be considered as a generalization of the (linear version of the) ‘Generative topographic mapping’ (Bishop et al, 1998), where the z_k form a fixed grid, and $\pi_k = 1/K$. Using a grid to capture the latent variable distribution may require a quite large value of K, especially when considering multivariate latent structures. Since our method recovers adaptively the latent variable distribution, the value K can be kept on a far smaller level (say, 6 or 7) even for a bivariate latent space (i.e., a plane). A further interesting aspect of the proposed technique is that, due to the generative model structure, it would allow additionally for inclusion of covariates in model (1). This would be attractive, for instance, for the computation of league tables from multivariate index data. This is a matter of further research.

Acknowledgments: The first author was supported by Nuffield Bursary URB/39353.00.

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 IWSM 1996*, 87–94, Orvieto, Italy.

Bishop, C.M J., Svensen, M. and Williams, C.K.I (1998) The Generative Topographic Mapping. *Neural Computation*, **10**, 215–234.

Bura, E. and Cook, R.D. (2001) Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society, Series B*, **63**, 393–410.