Measuring the component overlapping in mixtures of linear regressions

Susana Faria[1], Gilda Soromenho[2]

[1] Department of Mathematics and Applications, CMAT-Centre of Mathematics, University of Minho, Portugal;
[2] Institute of Education, University of Minho, Portugal

E-mail for correspondence: sfaria@math.uminho.pt

Abstract: Entropy-type measures for the heterogeneity of data have been used for a long time. In a mixture model context, entropy criterions can be used to measure the overlapping of the mixture components. In this paper we study an entropy-based criterion in mixtures of linear regressions to measure the closeness between the mixture components.
We show how an entropy criterion can be derived based on the Kullback-Leiber distance, which is a measure of distance between probability distributions. To investigate the effectiveness of the proposed criterion, a simulation study was performed.

Keywords: Mixtures of linear regressions; entropy criterion; Kullback-Leiber information; simulation study

1 Introduction

Finite mixture models are a well-known method for modelling data that arise from a heterogeneous population (e.g., see McLachlan and Peel, 2000; Fruhwirth-Schnatter, 2006 for a review). The study of these models is a well-established and active area of statistical research and mixtures of regressions have also been studied fairly extensively. Mixtures of linear regressions have also been studied extensively, especially when no information about membership of the points assigned to each line was available.
The mixture of linear regression model is given as follows:

$$y_i = \begin{cases} \mathrm{x}_i^T\beta_1 + \epsilon_{i1} & \text{with probability } \pi_1, \\ \mathrm{x}_i^T\beta_2 + \epsilon_{i2} & \text{with probability } \pi_2, \\ \qquad\vdots & \\ \mathrm{x}_i^T\beta_J + \epsilon_{iJ} & \text{with probability } \pi_J \end{cases} \tag{1}$$

where y_i is the value of the response variable in the ith observation; x_i^T $(i = 1, \ldots, n)$ denotes the transpose of the (p+1)-dimensional vector of indepen-

dent variables for the ith observation, β_j $(j = 1, \ldots, J)$ denotes the (p+1)-dimensional vector of regressor variables for the jth component, π_j are the mixing probabilities $(0 < \pi_j < 1$, for all $j = 1, \ldots, J$ and $\sum_j \pi_j = 1)$. Finally, ϵ_{ij} are the random errors; under the assumption of normality, we have $\epsilon_{ij} \sim N(0, \sigma_j^2)$, $(i = 1, \ldots, n; j = 1, \ldots, J)$.

2 Statistical Entropy

The Kullback-Leibler(KL) information (Kullback, 1959), also known as Kullback's directed divergence, is the measure of information discrepancy between $f(x)$ and $g(x)$, which is defined as

$$KL(f : g) = \int f(x) log \frac{f(x)}{g(x)} dx$$

where $f(x)$ is referred to as the reference distribution. The Kullback's directed divergence can be considered as a kind of a distance between the two probability densities, though it is not a real distance measure because it is not symmetric. An alternative directed divergence is the Kullback's symmetric divergence defined as the sum of two directed divergences (Frühwirth-Schnatter, 2006),

$$J(f : g) = KL(f : g) + KL(g : f) = \int f(x) log \frac{f(x)}{g(x)} dx + \int g(x) log \frac{g(x)}{f(x)} dx$$

(Leisch,2004) uses the Kullback-Leibler(KL) information to diagnose which components overlap in a mixture model.
Although there are many possible distance measures between two densities available in literature, the Kullback's symmetric divergence is attractive because of its simplicity and analytical tractability for mixtures models.

2.1 An entropy-based criterion

Consider a two-component mixture of linear regressions,

$$f(y|\mathbf{x}) = \pi_1 f_1(y; \mathbf{x}^T \beta_1, \sigma_1^2) + \pi_2 f_2(y; \mathbf{x^T} \beta_1, \sigma_2^2).$$

Based on Kullback's symmetric divergence, we define a criterion (EC) to study the overlapping of the mixture normal components in a two-component mixture of linear regressions,

$$\begin{aligned} EC(\pi_1 f_1 : \pi_2 f_2) &= KL(\pi_1 f_1 : \pi_2 f_2) + KL(\pi_2 f_2 : \pi_1 f_1) = \\ &= 2\pi_1 ln(\tfrac{\pi_1}{\pi_2}) + ln(\tfrac{\pi_2}{\pi_1}) + \pi_1 KL(f_1 : f_2) + \pi_2 KL(f_2 : f_1) \end{aligned}$$

with

$$KL(f_i : f_j) = \frac{n}{2}(ln \frac{\sigma_j^2}{\sigma_i^2} + \frac{\sigma_i^2}{\sigma_j^2}) + \frac{1}{2\sigma_j^2}(\mathbf{x^T} \beta_\mathbf{i} - \mathbf{x^T} \beta_\mathbf{j})^\mathbf{2} - \frac{\mathbf{n}}{\mathbf{2}}$$

3 Simulation study

To investigate the effectiveness of the proposed criterion, a simulation study was performed. We used the freeware R to develop the simulation program. Consider the mixtures of linear regressions,
mixture model 1: $f(y|\mathbf{x}) = \pi_1 f_1(y; \mathbf{x}^{\mathbf{T}}\beta_{f1}, \sigma^2_{f1}) + \pi_2 f_2(y; \mathbf{x}^{\mathbf{T}}\beta_{f2}, \sigma^2_{f2})$
mixture model 2: $g(y|\mathbf{x}) = \pi_1 g_1(y; \mathbf{x}^{\mathbf{T}}\beta_{g1}, \sigma^2_{g1}) + \pi_2 g_2(y; \mathbf{x}^{\mathbf{T}}\beta_{g2}, \sigma^2_{g2})$.
Samples of size $n = 100$ were generated for each set of true parameter values shown on Table 1 and the mixing proportion $\pi_1 = \{0.2, 0.4, 0.8\}$. We considered two typical configurations of the true regression lines: parallel and concurrent. For each type of simulated data set, 200 samples of size n were simulated.

TABLE 1. True parameter values for the essays

Configuration	Mixture 1						Mixture 2				
	β_{01}	β_{11}	β_{02}	β_{12}	σ^2_{f1}	σ^2_{f2}	β_{01}	β_{11}	β_{02}	β_{12}	σ^2_{f2}
concurrent/concurrent	2	1	4	-1	0.35	0.5	0	1	2	-1	0.5
paralell/paralell	1	1	3	1	0.35	0.5	3	1	5	1	0.5
paralell/concurrent	2	1	4	-1	0.35	0.5	3	1	5	1	0.5
concurrent/concurrent	0	1	2	-1	0.15	0.35	4	1	6	-1	0.5
paralell/paralell	1	1	4	1	0.3	0.35	3	1	6	1	0.5
paralell/concurrent	1	1	4	-1	0.3	0.35	2	1	4	-1	0.5

The simulation process consists of the following steps:

- Create a data set of size $n = 100$ of mixture model 1. Fit a mixture of linear regression models to the data using the EM algorithm. Save the estimated parameters $\hat{\Psi} = \{(\hat{\pi}_1, \hat{\pi}_2, \hat{\beta}_{f1}, \hat{\beta}_{f2}, \hat{\sigma}^2_{f1}, \hat{\sigma}^2_{f2})\}$ and calculate the estimated $\widehat{EC}$.
- Determine σ^2_{g1} so that the two mixture models (mixture model 1 and mixture model 2) have the same estimated EC value.
- Create a data set of size $n = 100$ of mixture model 2. Fit a mixture of linear regression models to the data using the EM algorithm. Save the estimated parameters $\hat{\Psi} = \{(\hat{\pi}_1, \hat{\pi}_2, \hat{\beta}_{g1}, \hat{\beta}_{g2}, \hat{\sigma}^2_{g1}, \hat{\sigma}^2_{g2})\}$ and calculate the estimated $\widehat{EC}$.
- Calculate the Mahalanobis distance between estimated and true parameters values in mixture model 1 and in mixture model 2.
- The Mann-Whitney test is used for testing the equality of the two Mahalanobis distances.

4 Conclusion

In this article, we define an entropy-based criterion in two component mixtures of linear regressions to measure the overlapping of the mixture components.
We note the following general findings:

- When the true regression lines are parallel in two mixtures models and the degree of overlap between two components is the same, there is no differences between the two estimates of the parameters of mixtures of linear regressions ;
- When the true regression lines are concurrent in two mixtures models and the degree of overlap between two components is the same, there is no differences between the two estimates of the parameters of mixtures of linear regressions ;
- When the true regression lines are parallel in one mixture model and concurrent in another mixture model and the degree of overlap between two components is the same, there is no differences between the two estimates of the parameters of mixtures of linear regressions.

We may conclude that the configurations of the true regression lines does not affect the performance of the EM algorithm, but only its degree of overlapping.

Acknowledgments: This research was financed by FEDER Funds through "Programa Operacional Factores de Competitividade COMPETE" and by Portuguese Funds through "FCT - Fundação para a Ciência e Tecnologia" , within the Project Est-C/MAT/UI0013/2011.

References

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models.* Springer, Heidelberg.

Kullback, S. (1959). *Information Theory and Statistics.* New York: Wiley.

Leisch, F. (2004). *Exploring the Structure of Mixture Model Components.* In J. Antoch (ed.),Compstat 2004 Proceedings in Computational Statistics, pp. 1405-1412. Physika Verlag, Heidelberg, Germany.

Mclachlan, G. J., and Peel, P. (2000). *Finite Mixture Models.* Wiley, New York.