First steps in the discovery of patterns in the academic results of telecommunication engineering students in the subjects Analysis of Circuits and Mathematics

Wilmar Hernandez
whernan@ics.upm.es
Universidad Politécnica de Madrid, Madrid,
Spain

Jorge Bonache
jbonache@euitt.upm.es
Universidad Politécnica de Madrid, Madrid
Spain

Abstract: *In this paper, the results of six years of research in engineering education, in the application of the European Higher Education Area (EHEA) to improve the performance of the students in the subject Analysis of Circuits of Telecommunication Engineering, are analysed taking into consideration the fact that there would be hidden variables that both separate students into subgroups and show the connection among several basic subjects such as Analysis of Circuits (AC) and Mathematics (Math). The discovery of these variables would help us to explain the characteristics of the students through the teaching and learning methodology, and would show that there are some characteristics that instructors do not take into account but that are of paramount importance.*

Introduction

The last six years have experienced a change in the teaching and learning methodology at university, in which professor of basic subjects of Telecommunication engineering such as Analysis of Circuits (AC) have moved from a traditional way of teaching to the application of the European Higher Education Area (EHEA) in order to improve the performance of the students.

The educational experiment presented in this paper was carried out in several stages. At the beginning, treatment and control groups were formed in AC, and some partial results were achieved. In the second stage, an analysis among all the subjects of first year students was carried out. The results of these two stages have been published in journals and international conferences on engineering education [Hernandez, Palmero et al. (2009); Hernandez, Palmero et al. (2010); Hernandez, Bonache et al. (2010)].

In [Hernandez, Bonache et al. (2010)], when conducting the statistical modelling of the student marks of AC, as a result, it was obtained that the observations were classified into two groups taking into consideration their probabilities of membership to these two groups . Hence, the threshold that divided both groups was found. Consequently, given the value x of X (mark), the threshold allowed us to decide which group an element belonged to. Now, we are interested to find the qualitative characteristics that distinguished the students of the groups that were obtained when classifying.

Taking into account the previous analysis for the marks, it can be said that there were some factors or hidden variables that were affecting significantly the academic results of the students and originating heterogeneity. That is to say, there were some factors that were producing a higher variability than the rest and, as a result, these factors were segmenting the population.

Here, it is important to point out that for the case under analysis the only information available was the marks of AC, which was not enough to look for the qualitative factors that differentiated the students of the two groups. However, it was suspected that one important factor, among others, was the following: how Mathematical tools are used in AC. The academic results in Math should have an influence in the ones in AC and could be one of the reasons for segmenting the population. Hence, in order to continue improving the performance of the students in AC, it could be interesting to study whether the results in Mathematics I (MATI) have influenced the ones in Analysis of Circuits I (ACI), and if the marks in MATI represent a factor that segment the mark in ACI.

Finite mixture of regression

In the third stage, which is the current one, the marks and trajectories of the students that were majoring in Sound Systems (SM) and Telecommunication Systems (TM) in the subjects ACI and Math, have been collected (MATISM, ACISM, MATITM, ACITM) for six years and compared with each other. Instructors of both subjects (i.e., ACI and MATI) have worked together and new interdisciplinary materials of study have been created [Hernandez (2010)].

The study that has been carried out from the marks of the students has been the following:

First, from the dispersion diagrams shown in Fig.1 and the calculation of the correlation coefficient (0.5345 for SM and 0.4822 for TM), it is observed both that there exists a linear relation between the independent variable MATISM (variable x) and the dependent variable ACISM (variable y), and that there also exists a linear relation between the independent variable MATITM (variable x) and the dependent variable ACITM (variable y). Hence, a linear regression model could be adjusted to the data. Nevertheless, when observing Fig. 1 deeply it can be seen that the marks in ACI are not homogeneous among students with similar marks in MATI. Therefore, it could be suggested that there are several groups for which a linear regression model would represent a good approximation. That is to say, there exists a different linear relation for groups of students for the level, the slope and the variability [Justel (2001)].

For the regression models it is assumed that the regression coefficients are the same for all the observations and it is also assumed that the sample (,)iiyx is a homogeneous group. In many cases, as it could be ours, the former assumption cannot be made if there are important variables that are not included in the model; that is, there is non-observed heterogeneity.

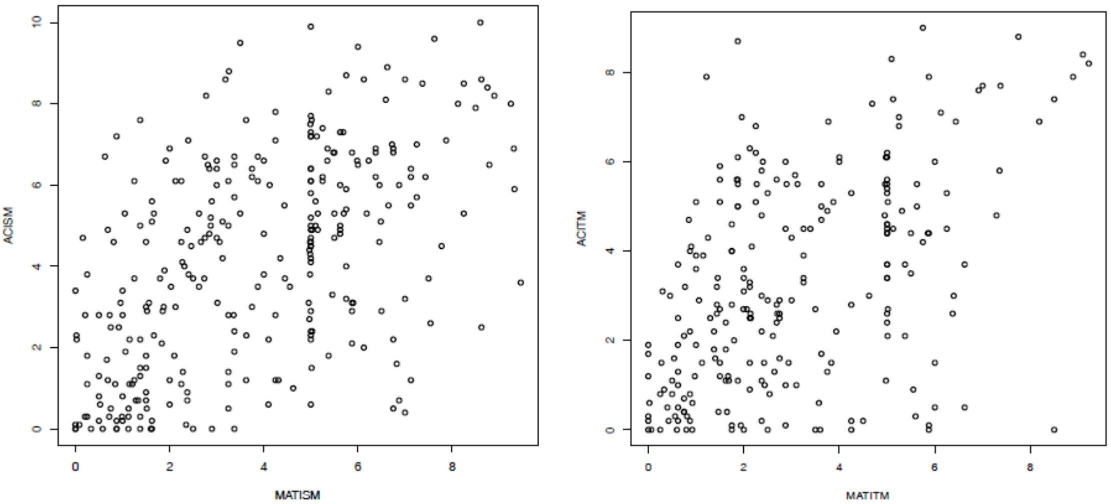


Figure 1: Dispersion diagrams: Marks of Math vs. ACI for students majoring in SM and TM

Therefore, second, the model that is going to be specified is a finite mixture of regression models. That is, a set of K regression models [Justel (2001); Hurn et al. (2003); Frühwith-Schnatter (2006); Bishop (2006); Leisch (2004)]

$$y = \alpha_i + \beta_i x + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_i^2)$$

characterized by their parameters:

$$(\alpha_1, \beta_1, \sigma_1^2), \dots\dots, (\alpha_K, \beta_K, \sigma_K^2)$$

Thus, the density function of the mixture is given by

$$f(y|x, \Psi) = \sum_{i=1}^{K} \pi_i \Phi_i\big(y|\alpha_i + \beta_i x, \sigma_i^2\big)$$

where $\Phi_i\big(\cdot\,|\alpha_i + \beta_i x, \sigma_i^2\big)$ represents the density function of the normal with mean $\alpha_i + \beta_i x$ and variance $\sigma_i^2, \Phi_i\big(\cdot\,|\alpha_i + \beta_i x, \sigma_i^2\big)$ was called a component of the mixture o class, π_i stands for *a priori* probabilities of such components, and Ψ stands for the set of all the parameters of the model.

Third, all the pairs (y_i, x_i) are observed and the following parameters are estimated:

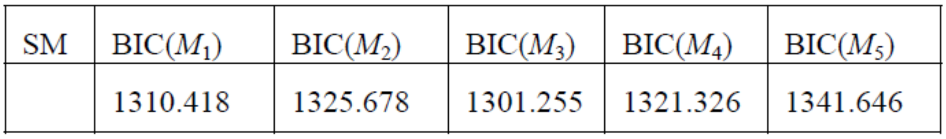
$$\pi_1, \dots, \pi_K, \qquad \alpha_1, \dots, \alpha_K, \qquad \beta_1, \dots, \beta_K, \qquad \sigma_1^2, \dots, \sigma_K^2, \qquad 0 < \pi_i < 1, \sum_{i=1}^{K} \pi_i = 1$$

In order to estimate the model parameters, the first step is to estimate the number of K-components of the model. To that end, the best model, M_i, will be chosen by using the Bayesian information criterion (BIC) [Hastie et al. (2001)]:

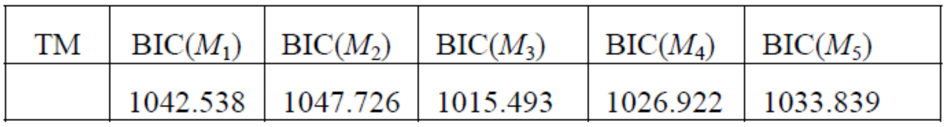
$$BIC_i = BIC(M_i) = -2 \log L(M_i) + p(M_i) \log n, \qquad i = 1, \dots, K$$

where L(M$_i$) represents the likelihood function for parameters in M$_i$, evaluated at the maximum likelihood estimators, and $p(M_i)$ represents the number of parameters of the model M_i. The model that will be chosen is the one with the smallest BIC.

Here, all the calculations are carried out by using the flexmix Package of R [R-project for statistical computing; Leisch (2004); Grün & Leisch (2007)], and the following results are shown:



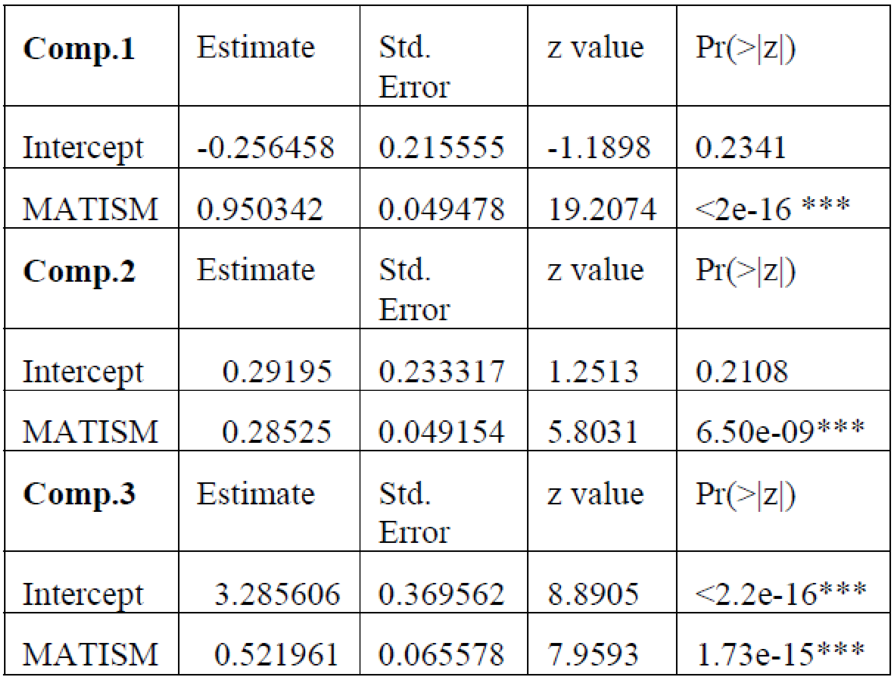
SM	BIC(M_1)	BIC(M_2)	BIC(M_3)	BIC(M_4)	BIC(M_5)
	1310.418	1325.678	1301.255	1321.326	1341.646



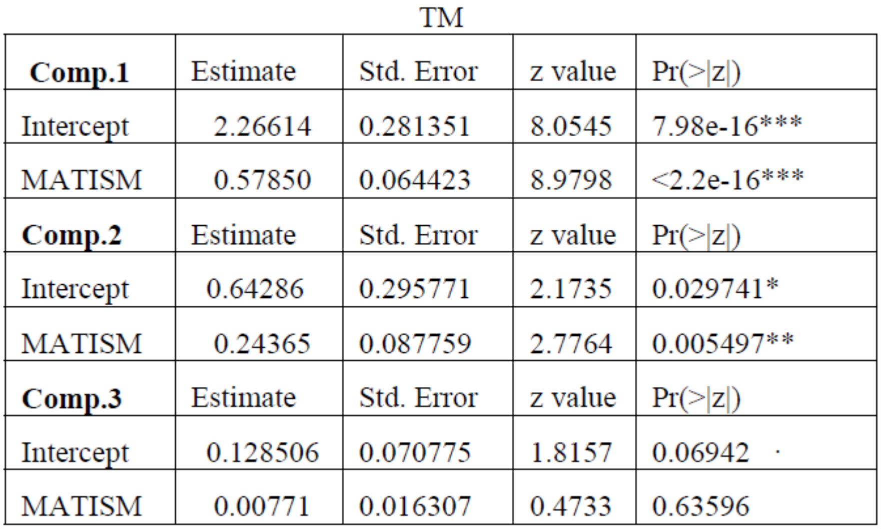
TM	BIC(M_1)	BIC(M_2)	BIC(M_3)	BIC(M_4)	BIC(M_5)
	1042.538	1047.726	1015.493	1026.922	1033.839

After K is determined (K = 3), the second step is the estimation of the parameters. To that end, the EM algorithm is used [Bishop (2006); Hastie et al. (2001)] and a test for significance of regression coefficient [Grün & Leisch (2007)] is carried out.

SM



Comp.1	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-0.256458	0.215555	-1.1898	0.2341
MATISM	0.950342	0.049478	19.2074	<2e-16 ***
Comp.2	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.29195	0.233317	1.2513	0.2108
MATISM	0.28525	0.049154	5.8031	6.50e-09***
Comp.3	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	3.285606	0.369562	8.8905	<2.2e-16***
MATISM	0.521961	0.065578	7.9593	1.73e-15***



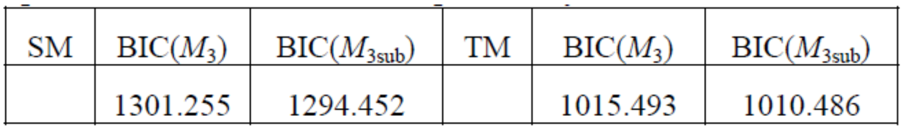
TM

Comp.1	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	2.26614	0.281351	8.0545	7.98e-16***
MATISM	0.57850	0.064423	8.9798	<2.2e-16***
Comp.2	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.64286	0.295771	2.1735	0.029741*
MATISM	0.24365	0.087759	2.7764	0.005497**
Comp.3	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.128506	0.070775	1.8157	0.06942 ·
MATISM	0.00771	0.016307	0.4733	0.63596

Usual R convention:

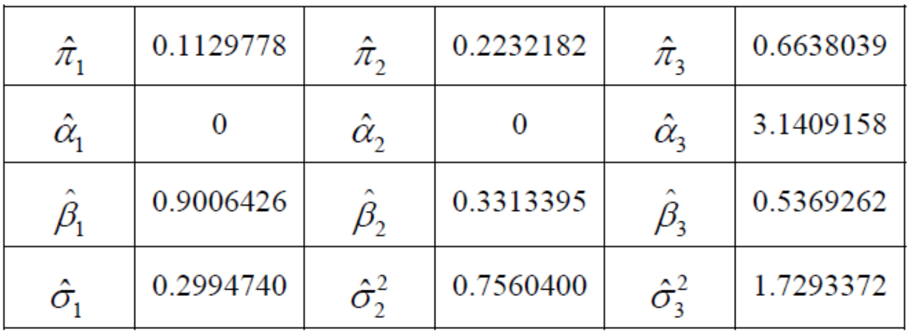
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 · 0.1 1

The above results, in both cases, tell us that know we should test a sub-model in which the ordinates in the origin of two components be equal to zero (for the case of SM) and the slope of the third component be equal to zero (for the case of TM). For SM and TM, the BIC of the sub-model is obtained and compared with the BIC that were previously obtained.



SM	BIC(M_3)	BIC($M_{3\text{sub}}$)	TM	BIC(M_3)	BIC($M_{3\text{sub}}$)
	1301.255	1294.452		1015.493	1010.486

Therefore, the sub-model for SM is chosen and the estimations of the parameters are the following:



$\hat{\pi}_1$	0.1129778	$\hat{\pi}_2$	0.2232182	$\hat{\pi}_3$	0.6638039
$\hat{\alpha}_1$	0	$\hat{\alpha}_2$	0	$\hat{\alpha}_3$	3.1409158
$\hat{\beta}_1$	0.9006426	$\hat{\beta}_2$	0.3313395	$\hat{\beta}_3$	0.5369262
$\hat{\sigma}_1$	0.2994740	$\hat{\sigma}_2^2$	0.7560400	$\hat{\sigma}_3^2$	1.7293372

In addition, the sub-model for TM is chosen and the estimations of the parameters are the following:

$\hat{\pi}_1$	0.6661254	$\hat{\pi}_2$	0.2202522	$\hat{\pi}_3$	0.1136224
$\hat{\alpha}_1$	2.4492977	$\hat{\alpha}_2$	0.4924507	$\hat{\alpha}_3$	0.1641922
$\hat{\beta}_1$	0.5536241	$\hat{\beta}_2$	0.5536241	$\hat{\beta}_3$	0
$\hat{\sigma}_1$	1.7428010	$\hat{\sigma}_1^2$	0.8606408	$\hat{\sigma}_1^2$	0.1751850

The estimated a posteriori probability that the i-th observation belongs to the j-th component 1, 2, 3 is given by [Bishop (2006); Leisch (2004); Hastie et al. (2001)]

$$\hat{P}(j/x,y,\hat{\Psi}) = \frac{\hat{\pi}_j \Phi_j(y/\hat{\alpha}_j + \hat{\beta}_j x, \hat{\sigma}_j^2)}{\sum_{i=1}^{3} \hat{\pi}_i \Phi_i(y/\hat{\alpha}_i + \hat{\beta}_i x, \hat{\sigma}_i^2)}$$

The *a posteriori* probabilities can be used to build groups or clusters with the data assigning each observation to the component with maximum *a posteriori* probability. For the case under analysis, the marks are classified in three groups (clusters) shown in Fig. 2.

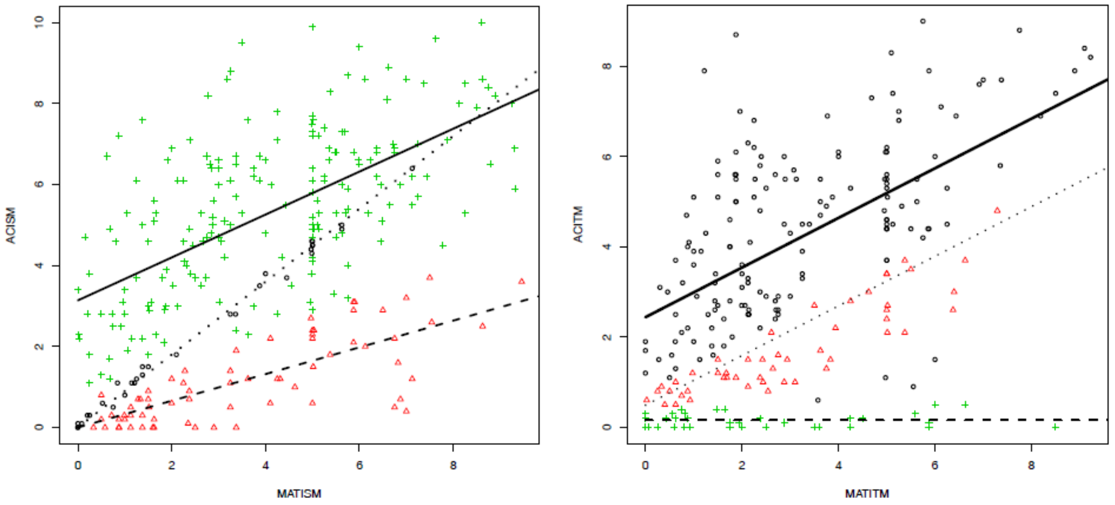


Figure 2: Clusters that are built with the data assigning each observation to the component with maximum a posteriori probability

SM	SM1	SM2	SM3	TM	TM 1	TM 2	TM 3
	(31) ○	(68) Δ	(195) +		(157) ○	(49) Δ	(35) +

$SM\ 1$	$y = 0.90x$	$TM\ 1$	$y = 2.44 + 0.55x$
$SM\ 2$	$y = 0.33x$	$TM\ 2$	$y = 0.49 + 0.55x$
$SM\ 3$	$y = 3.14 + 0.53x$	$TM\ 3$	$y = 0.16$

and the percentages of the classifications of each group are the following: SM1: 66.32%, SM2: 23.12%, SM3: 10.54%, TM1: 65.14%, TM2: 22.02%, TM3: 11.36%.

Conclusions

To sum up, for SM and TM there exist three heterogeneous groups in the data and we would have the attachment of each student to each one of these groups, as well. Moreover, it has been observed that in TM there is a small group (cluster) of students whose marks tend to oscillate around a constant mark in AC (TM3). In addition, it has been observed that also there are two groups (clusters) consisting of most of the students (approximately 2/3 parts) in which there would be a linear relation between the marks of AC and Math in SM and TM (SM3 and TM1).

Furthermore, observing TM1 and TM2, it could be suggested that there exists a hidden variable that will allow us to explain the membership to each one of these groups, which would occur in a similar but not so clear manner when observing SM2 and SM3. For the case of SM1, it results that the students have approximately the same marks in Math and AC. Hence, it should have to be taken into consideration the fact that there would be hidden variables that separate students into subgroups. The discovery of these variables would help us to explain the characteristics of the students through the teaching and learning methodology, and would show that there are some characteristics that instructors do not take into account but that are of paramount importance.

Future research plans

In order to discover significant differences among engineering students, it is important to study both the information collected from the students and the questions that would be interesting to ask them, because there would be hidden factors that we want to discover. To be more specific, from the finite mixture model that has been adjusted to the data, in order to try to find the variables that explain the above-mentioned clusters, we think that there is some information about the students that have not been taken into consideration yet. For instance, it is important to know the number of times each student has taken the subjects AC and Math. Furthermore, as these subjects are first-year, first-semester subjects, it could be possible that subjects such as Physics and Mathematics taught previously to entering university are influencing the performance of the students in their first academic year at university.

In addition, as during the educational experiment it was observed that there are groups whose marks oscillate around a constant value in the subject AC, this would suggest that these groups consist of students who have centered their preparation in Math and have probably abandoned AC. This abandonment could be analyzed by means of introducing the performance of the students during the course assignments of AC and Math, as another explanatory variable in the model.

Finally, it is important to point out that this educational experiment is an ongoing work, in which the next stage would consist of trying to discover the existence of specific characteristics of each cluster. To sum up, we would try to discover a pattern in the data, which obviously would mean an improvement in the teaching and learning process in AC and Math. To this end, besides of the marks of the students in AC and Math, for the next stage of the educational experiment it would be necessary to manage to get additional information from the students.

References

Hernandez, W., Palmero, J., Labrador, M., Alvarez-Vellisco, A., & Bonache, J. (2009). Analysis of Results of Application of a student-center Learning System to Improve Performance of First-year Students, *International Journal of Engineering Education,* 25(1), 161-172.

Hernandez, W., Palmero, J., Labrador, M., Bonache, J., Cousido, C., Alvarez-Vellisco, A., Gutierrez-Arriola, J. M., & Jimenez-Trillo, J. (2010). Analysis of the results of four years of research and application of a studentcentered system based on the ECTS to first-year students in order to improve their performance in the subject AC-I, *Proceedings of the 1st Annual Engineering Education Conference, IEEE EDUCON 2010* (pp. 237-242). Madrid, Spain: IEEE

Hernandez, W., Bonache, J., Cousido, C., Palmero, J., Labrador, M., & Alvarez-Vellisco, A. (2010). Statistical Analysis of Academic Results Before and After Four Years of Bologna. *International Journal of Engineering Education,* 26(6), 1493–1502.

Hernandez, W. (2010). *Herramientas Matemáticas para Análisis de Circuitos*. Madrid, Spain: Dpto. Publicaciones de la EUIT de Telecomunicación.

Justel, A. (2001). Estimación de mixturas de regresiones en el estudio de las emisiones de CO2 por países, *Proceedings of Conferencia Internacional de Estadística en Estudios Medioambientales, EMA'01* (pp. 107-113). Cádiz, Spain: Universidad de Cádiz.

Hurn, M., Justel, A., & Robert, C. P. (2003). Estimating Mixtures of Regressions. *Journal of Computational and Graphics Statistics,* 12(1), 55-79.

Frühwith-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models.* New York: Springer.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* New York: Springer.

Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11(8), 1-18.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Element of Statistical Learning*. New York: Springer. R-project for statistical computing. Accessed at www.r-project.org

Grün, B., & Leisch, F. (2007). Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis*, 51(11), 5247-5252.

Acknowledgements

This research has been supported by the Universidad Politécnica de Madrid under the research projects IE105902140, IE09590203, IE08590220 and IE075902067.

Copyright statement