A latent variable model for market segmentation

Francesca Greselin and Salvatore Ingrassia

Abstract We model a set of quantitative variables provided by a telecom company, and related to the amount of traffic of customers. Motivated by the high correlation observed among the variables, we employ mixtures of factor analyzers, which -at the same time- perform dimension reduction. The main purpose is to assess whether, within the same traffic plan, customers have a unique behavior in terms of traffic extent, or if different latent structures can be discovered. Any significative difference could be an important information for marketing, for instance to analyze patterns of pre-churn customers, or to identify specific targets. We implement a data-driven constrained approach for model estimation, to reduce spurious local maximizers and avoid singularities in the EM algorithm. First results highlight a non-unique latent structure for customers within the same traffic plan.

Key words: Market segmentation, Mixture of Factor Analyzers, Model-Based Clustering, Constrained EM algorithm.

1 Introduction and motivation

Finite mixture distributions are a very natural choice when it is assumed that a sample of observations arises from a specified number of underlying populations. Their central role is mainly due to their double nature: they combine the flexibility of non-parametric models with the strong and useful mathematical properties of parametric models.

Beyond these ”unconditional” approaches to finite mixtures of normal distributions, ”conditional” mixture models allow for the simultaneous probabilistic classification of observations and the estimation of regression models relating covariates to the expectations of the dependent variable within latent classes (see Wedel and De Sarbo, 1994, for a review). This methodology has been particularly employed in

Francesca Greselin
University of Milano Bicocca (Italy) e-mail: francesca.greselin@unimib.it

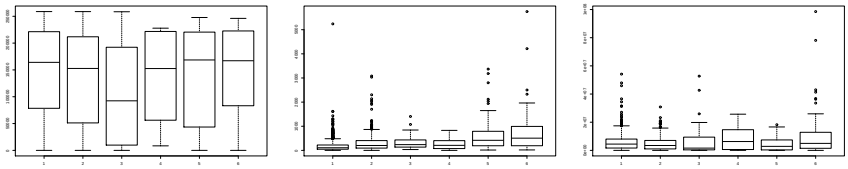
Salvatore Ingrassia
University of Catania (Italy)e-mail: s.ingrassia@unict.it

marketing research, due to the availability of categorical and ordinal data originated by surveys.

Our paper aims at modeling - through mixtures of Gaussian factor analyzers - a set of 64 quantitative variables provided by a telecom company, to estimate the latent structures among customers traffic. We will firstly resort to methodologies for variable selection (see, e.g., Liu *et al.*, 2003), as the non-informative variables can be strongly misleading for some clustering methods. Along the lines of Ghahramani and Hilton (1997) we then assume that the data have been generated by a linear factor model with latent variables modeled as Gaussian mixtures. Following Greselin and Ingrassia (2013), in this paper we maximize the likelihood function in a constrained parameter space, having no singularities and a reduced number of spurious local maxima.

2 Description of the dataset and explorative study

We deal with multivariate data provided by a telecom company, and related to the amount of traffic used by customers. Data concern a sample of 2172 customers, with 74 observed variables. Besides the personal data (ID, age, gender, city and province of residence) we have also more business-related informations on the customer (such as the type of handset owned, the aging as a client, whether the customer asked for number portability or not, the value of the customer in terms of rentability, etc.). Further, we have the complete record of the traffic extent, measured in terms of total usage (summed over 6 months: from August 2012 to January 2013), like: minutes of voice call, number of events of voice call, number of text messages, number of events of download and amount of downloaded data from the Internet. The data are divided into: ”traffic below the threshold of the plan” or ”out of it”, ”traffic On” or ”Off net”, and so on, summing up to 64 variables, which will be the focus of our analysis. Following Tukey’s approach, before modeling, we performed a first explorative study to see how each variable is distributed across the traffic plans, and some of our results are shown in Figure 1. To select the more important variables



(a) Min.Internet connection (b) Minutes to Voice Fixed (c) Downloaded Kb

Fig. 1 Summary for some traffic variables, from plan A to F (from left to right)

for the subsequent analysis, we adopted the random forest methodology, in the classification setting. This pre-step selects 7 final variables, with a loss of about 7.15% in terms of the Out-of-Box estimate of error rate, which increased from 16.55% to 23.7%. In particular, the classification error for units in plan A increased from 1.65% to 4.9%. The selected variables are the downloaded Kilobytes, the number of text messages sent Off and On net, the duration of Voice calls to Fixed line, or to

mobile phone, Off and On net. To allow for sensible handling through a Gaussian model, they were log-transformed (after adding a positive shift). Figure 2 shows the empirical distribution of the 7 selected log-variables (plus variable "duration of voice call to International line"), within traffic Plan A, selected by 1449 customers.

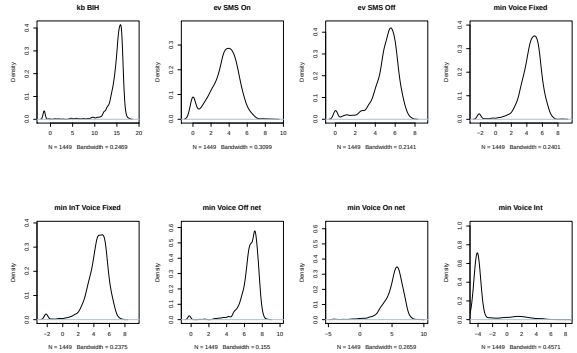


Fig. 2 Empirical distribution of the 7 selected log-variables in Plan A (kernel density estimated), plus variable Voice call to International

3 Do customers have a similar traffic extent?

In the following we confine the analysis to plan A, which is the crucial group of customers for the company, as it represents 69.93% of the clients. Our aim is to analyze if the marginal densities we observe in Figure 2 can be jointly modeled by a mixture of multivariate Gaussian factors, in such a way that different underlying behavior of customers, in terms of traffic usage, can be assessed for further market analyses. The constrained estimation has been implemented by a data-driven method, which will be briefly described below (for further details, see Greselin and Ingrassia 2013). Let $\mathbf{X}$ denote the matrix of the data, we firstly run the constrained algorithm with upper bound taking values in a set of k values, ranging from $\lambda^* = \lambda_{\max}(\mathrm{Cov}(\mathbf{X})) = 11.56587$ to $\lambda_* = \lambda_{\min}(\mathrm{Cov}(\mathbf{X})) = 0.005496$, and stopping when we get a decrease in the final likelihood. The grid of values is chosen in such a way that the spacings (intervals between subsequent values) follow a geometric series, i.e. giving a steady relative increase in the bounds. The suitable upper bound U is hence selected as the last bound which produced an improvement on the final likelihood. An analogous procedure for the selection of the lower bound is then applied. After choosing an ε value to protect from singularities, a grid of k values from ε to λ_* is employed for the lower bound in the constrained algorithm (with fixed upper bound U), and stopping when we observe a decrease in the final likelihood, picking the second last value as suitable lower bound L.

Table 1 Grid of values for bounds in constrained Mixtures of FA on Plan A (with $k = 5$, $\varepsilon = 10^{-6}$)

Upper bound grid	0.319864	0.634231	1.262966	2.520437	5.035378	10.065260
Lower bound grid	0.000181	0.000353	0.000696	0.001382	0.002753	0.005496

Now, to compare our proposal with similar approaches, the direct benchmarks in the literature are mixtures of Common Factor Analyzers (*Mcfa*) (McLachlan *et al.*, 2003) and Parsimonious Gaussian Mixtures of Factor Analyzers (*Pgmm*). We further take into account Gaussian mixtures, estimated through *Mclust* package in *R*. The latter provides three best fitting solutions: a mixture of $G = 9$ components, VEV structure for the covariance, with $BIC = 22198.9$; a VEV model with $G = 10$ and $BIC = 22281.2$; and finally a VEV model with $G = 7$ and $BIC = 22342.2$. We recall that $BIC = -2\log\mathscr{L} + k\log(n)$ is a penalized likelihood criterion, where n is the sample size and k is the number of estimated parameters. *Mcfa* results provide, considering the number of groups G ranging from 1 to 8 and the dimension of the latent factors q varying from 1 to 6, that the best model has $G = 7$ and $q = 6$, with $BIC = 20123.3$. On the other hand, *Pgmm* best model has $G = 4$ and $q = 3$, with $BIC = 20157.9$ and unconstrained covariances. All methodologies discard the existence of only one group among observations (for $G = 1$, $BIC = 27226.4, 27230.7$ and 27201.6 respectively for *Mcfa*, *Pgmm* with $q = 5$, and *Mclust*). Our method provides the overall best fit, with $G = 6$ groups, latent factors of dimension $q = 5$, and $BIC = 16208.8$, and further work is needed to interpret the latent structure. Using lighter constraints, derived by the data, our proposal allows for a better fit to the data, while the constraints adopted by the comparing models are too strong for the dataset at hand.

Table 2 *BIC* results for models in the family of Mixtures of Factor Analyzers, on Plan A (max 50 iter, max 10 init), for some values of the latent dimension q and number of groups G

	Mcfa				*Constrained Mixtures of FA*			
$q \setminus G$	*4*	*5*	*6*	*7*	*4*	*5*	*6*	*7*
4	25200.2	25158.8	24607.8	24683.4	18955.2	18058.7	18312.7	17620.0
5	22580.8	23510.5	21228.3	21072.9	17559.4	16866.1	**16208.8**	16553.5
6	21722.8	21494.8	20497.0	**20123.3**	17721.2	18206.8	17098.8	17010.5

References

Ghahramani, Z. and Hilton, G. (1997). The EM algorithm for mixture of factor analyzers. *Techical Report CRG-TR-96-1*.

Greselin, F. and Ingrassia, S. (2013). Maximum likelihood estimation in constrained parameter spaces for mixtures of factor analyzers. *Statistics and Computing*, pages DOI: 10.1007/s11222–013–9427–z, forthcoming.

Liu, J., Zhang, J., Palumbo, M., and Lawrence, C. (2003). Bayesian clustering with variable and transformation selection (with discussion). *Bayesian Statistics*, **7**, 249–275.

McLachlan, G., Peel, D., and Bean, R. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis*, **41**, 379–388.

Wedel, M. and De Sarbo, W. (1994). A review of recent developments in latent class regression models. In R. Bagozzi, editor, *Advanced Methods of Marketing Research*, pages 352–388. Blackwell, Cambridge, MA.