Data driven constraints for Gaussian mixtures of factor analyzers: an application to market segmentation

Francesca Greselin

Francesca Greselin - Università di Milano-Bicocca,
Dipartimento di Statistica e Metodi Quantitativi
francesca.greselin@unimib.it

Salvatore Ingrassia - Università di Catania,
Dipartimento di Economia e Impresa
s.ingrassia@unict.it

November 8, 2013

AG DANK/BCS Meeting 2013 - University College London

Outline

- We want to make a first explorative analysis on traffic usage for a telecom company
- Mixtures of factor analyzers, estimated through EM
- BUT: maximization of the log-likelihood without any constraint is an ill-posed problem (Day, 1969)
- to reduce spurious local maximizers and avoid singularities, some authors propose to take a common (diagonal) error matrix (MCFA Baek *et al.*, 2010) or to impose an isotropic error matrix (Bishop and Tippin, 1998)
- our proposal: a less constrained approach, based on covariance decomposition
- a first application is shown, suggesting a non-unique behavior of customers inside the traffic plan

Methodology and Aim

Our proposal is to adopt a weakly constrained approach for ML estimation,

- having no singularities, and
- simultaneously reducing the number of spurious local maxima

Aim

Provide market segmentation for telecom data, by using a latent variable approach, based on constrained mixtures of gaussian factor analyzers

The data

A sample of 2072 customers (postpaid plans)
with 45 quantitative variables about traffic usage (tot over 6 mths: Aug'12-Jan'13), like

- minutes of voice call (Off net, On net, International, to Fixed line)
- number of events of voice call (Off net, On net, Int, to Fix. l.)
- number of sent SMS (Off net, On net)
- number of events of data download from Internet
- amount of downloaded data (in Kb)
- minutes of data download
- number of events of data download in roaming or GPRS
- amount of downloaded data in roaming or GPRS (in Kb)
- minutes of data download in roaming or GPRS

Data is divided into:
total / under / over the threshold of the plan / no threshold

Further, we have 10 qualitative variables (ID, age, sex, geographic location (2 var), aging as a customer, value, price plan, handset, portability)

One of the (many) open questions in the market

When the market is saturated, the pool of available customers is limited and an operator has to shift from its acquisition strategy to retention because the cost of acquisition is typically five times higher than retention.

As noted in (Mattersion, 2001)
For many telecom executives, figuring out how to deal with Churn is turning out to be the key to very survival of their organizations.

Based on marketing research (Berson *et al.*, 2000), the average churn of a wireless operator is about 2% per month. That is, a carrier looses about a quarter of its customer base each year.

We need a model to understand the data and to devise patterns of pre-churn customers.

A first step: Exploratory Data Analysis

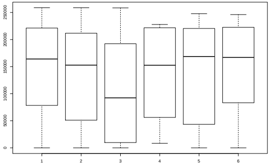
EDA is an approach to analyzing data sets to summarize their main characteristics (Hoaglin *et al.*, 2000), opening some questions in our minds

- what the data can tell us
- what assumptions could be reasonably be made w.r.t. the actual data
- what kind of model could be fit
- what set of hypotheses could be assessed
- ...

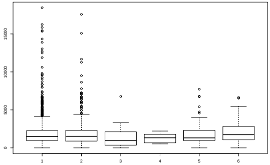
Some questions:

- Is the traffic usage highly related to the traffic plan?
- Which variables are more related to the customer experience?
- Does the plan affect the mean duration of the call? Or the mean amount of download? Or the mean number of SMS?
- Is the customer experience influenced by the part of the plan he does not exploit?
- Is it possible to identify pre-churn customers?
- How could the company be aware about new customers needs?
- How could the company propose a customer his plan?

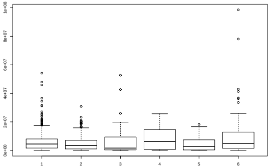
Tukey promoted the use of a five number summary for quantitative data:



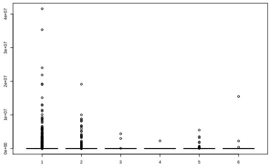
(a) Minutes



(b) No. of events



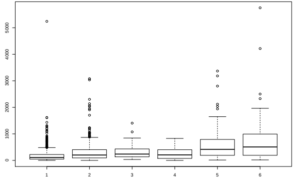
(c) Kbytes



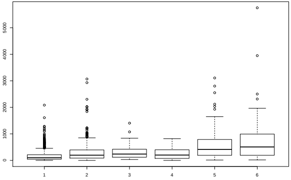
(d) Kbytes ExThs

Figure: Summary for Big Internet Home Data, from plan A to F

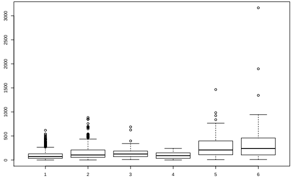
Exploratory Data Analysis 2



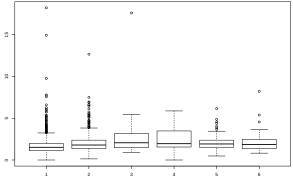
(a) Minutes



(b) Minutes InThs

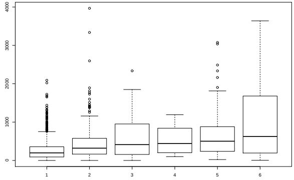


(c) No. of phone calls

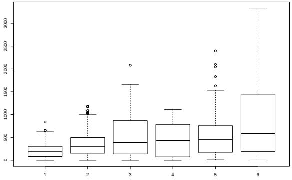


(d) Mean duration ExThs

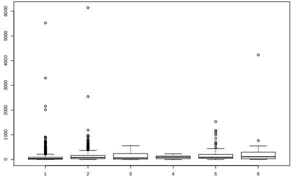
Figure: Summary for VOICE to Fixed calls, from plan A to F



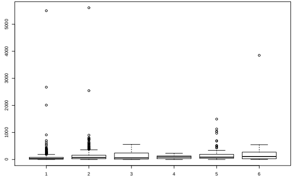
(a) No. of SMS Off net



(b) No. of SMS Off net InT



(c) No. of SMS On net



(d) No. of SMS On net InT

Figure: Summary for number of SMS sent Off (upper row) and On Net

Exploratory Data Analysis 4

How is the ”number of phone calls” variable distributed into the different plans?

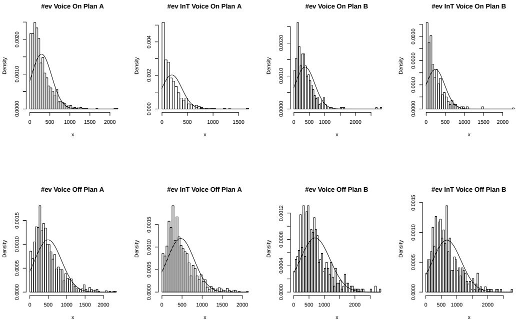


Figure: No. of phone calls On and Off Net - plan A (left) and B (right)

Exploratory Data Analysis 5

Are mean values ”better” distributed than original variables?

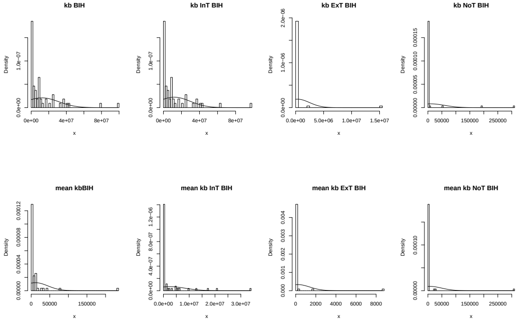


Figure: Kb and Mean Kb downloaded via Internet - plan F

Variable selection

To select the more important variables
we adopted the random forest methodology:

Type of random forest: classification
Number of trees: 10000
No. of variables tried at each split: 7

We pass from the 45+20 (original+mean values) to 7 final variables, by steps, each time deleting the 10 less important variables

the OOB estimate of error rate increases from 16.55% to 19.79%

The 7 selected variables

- Kb BIH
- ev SMS On
- evSMS Off
- min Voice to Fixed
- min InT Voice to Fixed
- min Voice Off net
- min Voice On net

Empirical distribution densities

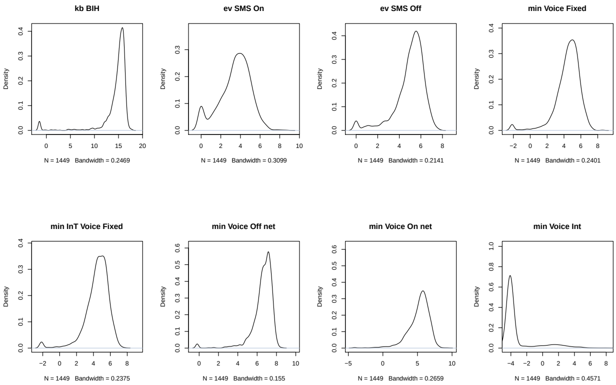


Figure: Empirical distribution of the 8 log transformed variables in Plan A (kernel density estimated), sample of 1449 units

Empirical distribution densities

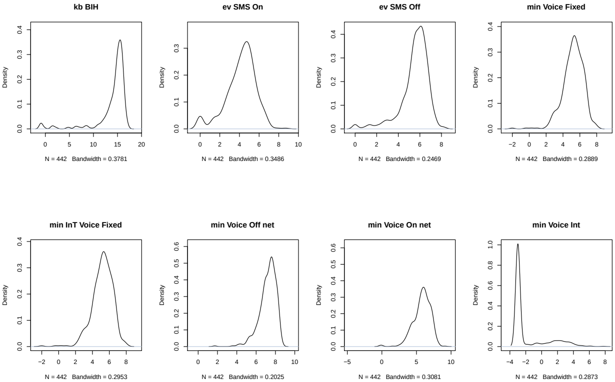


Figure: Empirical distribution of the 8 log transformed variables in Plan B (kernel density estimated), sample of 442 units

Empirical distribution densities

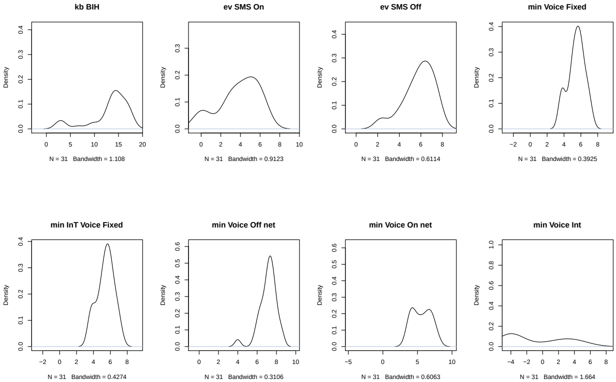


Figure: Empirical distribution of the 8 log transformed variables in Plan C (kernel density estimated), sample of 31 units

Empirical distribution densities

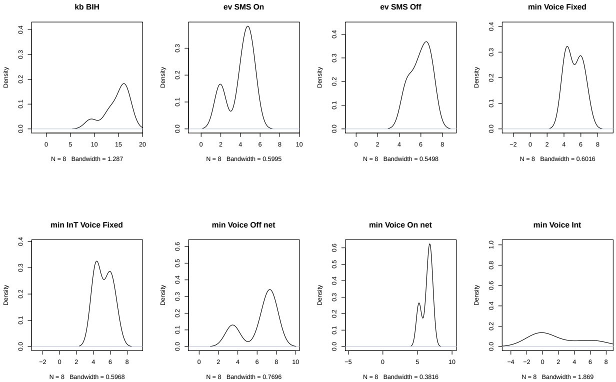


Figure: Empirical distribution of the 8 log transformed variables in Plan D (kernel density estimated), sample of 8 units

Empirical distribution densities

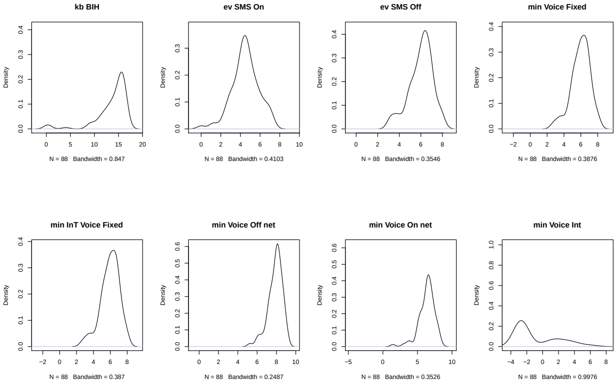


Figure: Empirical distribution of the 8 log transformed variables i in Plan E (kernel density estimated), sample of 88 units

Empirical distribution densities

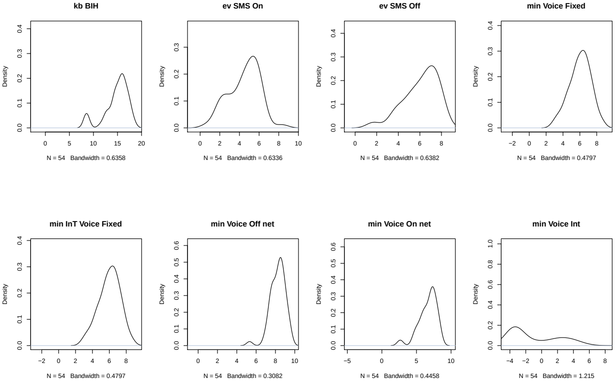


Figure: Empirical distribution of the 8 log transformed variables in Plan F (kernel density estimated), sample of 54 units

Mixture of Gaussian Factor Analyzers

Let $f(\mathbf{x};\boldsymbol{\theta})$ be the density of the d-dimensional random variable $\mathbf{X}$

$$f(\mathbf{x};\boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g \phi_d(\mathbf{x};\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$$

MGFA explain the correlation between a set of d variables in terms of a lower number q of underlying factors:

$$\mathbf{X}_i = \mu_g + \boldsymbol{\Lambda}_g \mathbf{U}_{ig} + \mathbf{e}_{ig} \quad \text{with prob} \quad \pi_g \quad \text{for } i = 1, \dots, n, g = 1, \dots, G$$

where

$\boldsymbol{\Lambda}_g$ is a $d \times q$ matrix of factor loadings,
$\mathbf{U}_{1g}, \dots, \mathbf{U}_{ng} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q)$ are the factors, ind. w.r.t. $\mathbf{e}_{ig}$,
$\mathbf{e}_{ig} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_g)$ are the errors with $\boldsymbol{\Psi}_g$ $d \times d$ diagonal matrix.

Under these assumptions,

$$\mathbf{\Sigma}_g = \mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \mathbf{\Psi}_g, \qquad d(q+1) \quad \text{params.}$$

and the parameter vector is

$$\theta_{GMFA}(d, q, G) = \{\boldsymbol{\mu}_g, \mathbf{\Lambda}_g, \mathbf{\Psi}_g, \pi_g (g = 1, \ldots, G-1)\}$$

The EM algorithm for MGFA

Given an initial random clustering $\mathbf{z}^{(0)}$, on the $(k+1)-th$ iteration,

1. Compute $z_{ig}^{(k+1)}$ and consequently obtain $\pi_g^{(k+1)}$ and $\boldsymbol{\mu}_g^{(k+1)}$ and also $n_g^{(k+1)}$ and $\mathbf{S}_g^{(k+1)}$ in the usual way;
2. Set a starting value for $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$ from $\mathbf{S}_g^{(k+1)}$;
3. Iterate the following steps, until convergence on $\hat{\boldsymbol{\Lambda}}_g$ and $\hat{\boldsymbol{\Psi}}_g$:
   1. $\gamma_g \leftarrow \gamma_g^+ = \boldsymbol{\Lambda}_g^{'}(\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g^{'} + \boldsymbol{\Psi}_g)^{-1}$ and
      $\boldsymbol{\Theta}_g \leftarrow \boldsymbol{\Theta}_g^+ = \mathbf{I}_q - \gamma_g\boldsymbol{\Lambda}_g + \gamma_g\mathbf{S}_g^{(k+1)}\gamma_g^{'}$;
   2. $\boldsymbol{\Lambda}_g \leftarrow \boldsymbol{\Lambda}_g^+ = \mathbf{S}_g^{(k+1)}\gamma_g^{'}(\boldsymbol{\Theta}_g^{-1})$ and
      $\boldsymbol{\Psi}_g \leftarrow \boldsymbol{\Psi}_g^+ = \text{diag}\left\{\mathbf{S}_g^{(k+1)} - \boldsymbol{\Lambda}_g^+\gamma_g\mathbf{S}_g^{(k+1)}\right\}$;
4. Compute $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$ and evaluate the log-likelihood, to check for convergence.

ML in constrained parametric spaces

The maximization of $\mathcal{L}$ over $\theta_{GMFA}(d, q, G)$ is an ill-posed problem. Further, a number of spurious maximizers could arise.

Hathaway (1985) proposed a constrained ML: Let $c \in (0, 1]$, then the following constraints

$$\min_{1 \leq h \neq j \leq k} \lambda(\mathbf{\Sigma}_h \mathbf{\Sigma}_j^{-1}) \geq c \tag{1}$$

on the eigenvalues λ of $\mathbf{\Sigma}_h \mathbf{\Sigma}_j^{-1}$ leads to properly defined, scale-equivariant, consistent ML-estimators for the mixture-of-normal case.

ML in constrained parametric spaces

To assure (1) we can impose the stronger condition

$$a \leq \lambda_{ig} \leq b, \qquad i = 1, \dots, d; \qquad g = 1, \dots, G \tag{2}$$

where $\lambda_{ig} = \lambda_i(\mathbf{\Sigma}_g)$, and $a, b \in \mathbb{R}^+$: $a/b \geq c$, see Ingrassia (2004).

Due to the structure of the covariance matrix $\mathbf{\Sigma}_g$, (2) translates into

$$a \leq \lambda_{ig}(\mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \mathbf{\Psi}_g) \leq b$$

ML in constrained parametric spaces

Finally, we set

$$
\begin{aligned}
d_{ig}^2 + \psi_{ig} &\geq a & i &= 1, \ldots, d & (3)\\
d_{ig} &\leq \sqrt{b - \psi_{ig}} & i &= 1, \ldots, q & (4)\\
\psi_{ig} &\leq b & i &= q+1, \ldots, d & (5)
\end{aligned}
$$

for $g = 1, \ldots, G$, where d_{ig} denote the singular values of $\mathbf{\Lambda}_g$ and ψ_{ig} denote the eigenvalues of $\mathbf{\Psi}_g$.

In particular, (3) reduces to $\psi_{ig} \geq a$ for $i = (q+1), \ldots, d$.

How do we choose constraints?

If we do not have any a priori information on a, b, or c, choosing the constrained parameter space is a difficult issue.

The constant c can be chosen by computing the profile $\mathcal{L}(c)$, for some set of grid points $c \in (0, 1]$ (Yao, 2010)

Rocci (2012) compute c by cross validation.

Both methods are computationally intensive.

We expect that the constrained algorithm, run with different values of the bounds, can give us a hint on how to choose them properly, by observing the final $\mathcal{L}(c)$.
Optimal values of the bounds should correspond to some agreement, over different random starts, on optimal values of $\mathcal{L}(c)$. Conversely, a simultaneous drop in $\mathcal{L}(c)$ observed for a new bound, over different random starts, indicates that the new constraint is too strong for the data at hand.

Data-driven upper bound

Procedure: *Choice of the upper bound b*

1. compute $\mathrm{Cov}(\mathcal{S})$ of sample $\mathcal{S}$ and set $\lambda^* = \lambda_{\max}(\mathrm{Cov}(\mathcal{S}))$;
2. choose an integer m and set $\mathbf{b} = (b_1, \ldots, b_m) \in \mathbb{R}^m$ where

$$b_j = \frac{j}{m}\lambda^* \qquad \text{for } j = 1, \ldots, (m-1) \qquad b_m = +\infty;$$

3. for $j = m, m-1$ run the unconstrained EM algorithm with $b = b_j$ and evaluate $\mathcal{L}_j$;
4. while $j > 1$ and $\mathcal{L}_j \geq \mathcal{L}_{j+1}$:
   - decrease j;
   - run the constrained EM algorithm with $b = b_j$ and evaluate $\mathcal{L}_j$;
5. set $b = b_{j-1}$ and $\hat{\boldsymbol{\theta}} = \arg_\theta \max \mathcal{L}_{j-1}(\boldsymbol{\theta})$.

An analogous procedure can be devised for the lower bound, after setting $\lambda_* = \lambda_{\min}(\mathrm{Cov}(\mathcal{S}))$, for more details see Greselin and Ingrassia (2013).

Mixtures of Factor Analyzers with Common Factor Loadings

We want to compare our proposal with the well known MCFA model. The latter is a recent method to deal with "constrained" maximization for EM, which at the same time allows for greater reduction in the number of parameters. The authors add the two following constraints (Baek *et al.*, 2010)

$$\mu_g = \mathbf{A}\xi_g$$

and

$$\mathbf{\Sigma}_g = \mathbf{A}\mathbf{\Omega}_g\mathbf{A}' + D$$

Application: How is traffic usage in Plan A?

Table: Results of constrained GMFA on Plan A (69.93% of customers) for $d = 7, q = 4, G = 2$

run	No iter	$\mathcal{L}_0$	$\mathcal{L}_{fin}$	BIC	α_1	α_2
0	71	-12337.53	-11404.84	23428.37	0.6541167	0.3458833
1	60	-13843.82	-11404.85	23428.39	0.6541363	0.3458637
2	16	-13726.44	-10859.68	22338.05	0.7537929	0.2462071
3	34	-13681.53	-10859.69	22338.06	0.7537701	0.2462299
4	59	-13716.74	-11404.85	23428.38	0.6541293	0.3458707
5	53	-13926.21	-11404.85	23428.38	0.6541314	0.3458686
6	41	-13712.32	-11404.85	23428.38	0.3458635	0.6541365
7	7	-13839.80	-11594.34	23807.36	0.7654098	0.2345902
8	39	-13810.36	-10859.69	22338.06	0.2462301	0.7537699
9	30	-13824.78	-10859.69	22338.07	0.2462284	0.7537716
10	40	-13948.37	-10859.69	22338.06	0.7537700	0.2462300
11	34	-13493.07	-10859.69	22338.07	0.2462288	0.7537712
12	35	-13714.11	-11112.42	22843.51	0.7260618	0.2739382
13	35	-13864.51	-10859.69	22338.06	0.2462299	0.7537701
14	34	-13827.36	-10859.69	22338.06	0.2462297	0.7537703
15	54	-13802.20	-11404.84	23428.37	0.6541133	0.3458867
16	28	-13686.70	-11588.50	23795.69	0.7827988	0.2172012
17	45	-13862.32	-10859.69	22338.06	0.2462295	0.7537705
18	35	-13697.06	-10859.69	22338.07	0.2462280	0.7537720
19	34	-13883.69	-10859.69	22338.07	0.7537716	0.2462284
20	21	-13788.65	-11394.43	23407.55	0.3971774	0.6028226

Table: Vector **b** of values for the upper bound in constrained ML $\lambda^* = 11.56587$

2.313174	4.626348	6.939522	9.252696	11.56587	∞

Application: How is traffic usage in Plan A?

Table: Results of MCFA on Plan A (max 50 iter, max 50 init)

d	$\mathcal{L}_{fin}$	BIC
2	-13664	27532
3	-12917	26111
4	-12573	25496
5	-12684	24789
6	-12666	24828

where BIC= $-2\log\mathcal{L}_{fin} - k\log(n)$,
n is the sample size and k is the number of estimated parameters.

The Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models, based in a penalized log-likelihood. The best model is the one with lower BIC.

Conclusions

Aiming at modeling traffic usage, we have employed mixtures of gaussian factors analyzers.

To face the estimation issues, we considered a constrained approach, where the bounds can be obtained by a data-driven method.

We compared our results to the well known MCFA approach on the largest subsample of customers.

First results reveal at least two different behaviors among the customers, even inside the same plan.

References

Baek, J., McLachlan, G., and Flack, L. (2010). Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualization of high-dimensional data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **32**(7), 1298 –1309.

Berson, A., Smith, S., and Thearling, K. (2000). Building data mining applications for crm. *New York*.

Bishop, C. M. and Tippin, M. E. (1998). A hierarchical latent variable model for data visualization. *IEEE Transactions on Pattern analysis and Machine Intelligence*, **20**, 281–293.

Day, N. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, **56**(3), 463–474.

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

Hathaway, R. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, **13**(2), 795–800.

Hoaglin, D. C., Mosteller, F., and Tuckey, J. W. (2000). *Understanding Robust and Exploratory Data Analysis*. Wiley Classic Library Edition, NewYork.

Ingrassia, S. (2004). A likelihood-based constrained algorithm for multivariate normal mixture models. *Statistical Methods & Applications*, **13**, 151–166.

Mattersion, R. (2001). *Telecom Churn Management*. APDG Publishing, Fuquay-Varina, NC.

Rocci, R. (2012). Gaussian mixture models: constrained and penalized approaches. In *MBC*^{2} *Workshop, Catania (Italy)*.

Yao, W. (2010). A profile likelihood method for normal mixture with unequal variance. *Journal of Statistical Planning and Inference*, **140**(7), 2089 – 2098.