Variable selection in model-based classification

G. Celeux[1], M.-L. Martin-Magniette[2], C. Maugis[3]

1: INRIA Saclay-Île-de-France
2: UMR AgroParisTech/INRA MIA 518 et URGV (Unité de Recherche en Génomique Végétale)
3: Institut de Mathématiques de Toulouse

Variable selection in clustering and classification

- Variable selection is highly desirable for unsupervised or supervised classification in high dimension setttings.
- Actually, this question received a lot of attention in recent years.
- Different variable selection procedures have been proposed from heuristic point of views.
- Roughly speaking, the variables are separated into two groups : the relevant variables and the independent variables.
- In the same spirit, sparse classification methods have been proposed depending on some tuning parameters.
- We opt for a mixture model which allows to deal properly with variable selection in classification.

Gaussian mixture model for clustering

- Purpose : Clustering of $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ where $\mathbf{y}_i \in \mathbb{R}^Q$ are iid observations with unknown pdf h

- The pdf h is modelled with a Gaussian mixture

$$f_{\text{clust}}(.|K, m, \alpha) = \sum_{k=1}^{K} p_k \Phi(.|\mu_k, \Sigma_k)$$

with

- $\alpha = (\mathbf{p}, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$ where $\mathbf{p} = (p_1, \ldots, p_K)$, $\sum_{k=1}^{K} p_k = 1$
- $\Phi(.|\mu_k, \Sigma_k)$ the pdf of a $\mathcal{N}_Q(\mu_k, \Sigma_k)$

- $\mathcal{T}$ = set of models (K, m) where
  - $K \in \mathbb{N}^\star$ = number of mixture components
  - m = Gaussian mixture type

The Gaussian mixture collection

- It is based on the eigenvalue decomposition of the mixture component variance matrices :

$$\Sigma_k = L_k D_k' A_k D_k$$

  - Σ_k variance matrix with dimension $Q \times Q$
  - $L_k = |\Sigma_k|^{1/Q}$ (cluster volume)
  - $D_k = \Sigma_k$ eigenvector matrix (cluster orientation)
  - $A_k = \Sigma_k$ normalised eigenvalue diagonal matrix (cluster shape)

- $\Rightarrow$ 3 families :

$$\left.\begin{array}{l}\text{spherical family}\\ \text{diagonal family}\\ \text{general family}\end{array}\right\} \Rightarrow 14 \text{ models}$$

- Free or fixed proportions

- $\Rightarrow$ 28 Gaussian mixture models

Model selection

- Asymptotic approximation of the integrated or completed integrated likelihood
- BIC (Bayesian Information Criterion)

$$2\ln\left[f\left(\mathbf{y}|K,m\right)\right] \approx 2\ln[f(\mathbf{y}|K,m,\hat{\alpha})] - \lambda_{(\mathrm{K,m})}\ln(n) = \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}|K,m)$$

where $\hat{\alpha}$ is computed by the EM algorithm.

- ICL (Integrated Likelihood Criterion)
  ICL $=$ BIC $+$ Entropy of the fuzzy clustering matrix.

- The classifier : $\hat{z} = \mathrm{MAP}(\hat{\alpha})$ is

$$\hat{z}_{ik} = \begin{cases} 1 & \text{if } \hat{p}_k\Phi(\mathbf{y}_i|\hat{\mu}_k,\hat{\Sigma}_k) > \hat{p}_j\Phi(\mathbf{y}_i|\hat{\mu}_j,\hat{\Sigma}_j), \forall j \neq k \\ 0 & \text{otherwise} \end{cases}$$

MIXMOD software

http ://www.mixmod.org

Variable selection in the mixture setting

- **Law, Figueiredo and Jain (2004)** :
  The irrelevant variables are assumed to be independent of the relevant variables.

- **Raftery and Dean (2006)** :
  The irrelevant variable are linked with all the relevant variables according to a linear regression.

- **Maugis, Celeux and Martin-Magniette (2009a, b)** :
  SRUW Model
  The irrelevant variables could be linked to a subset of the relevant variables according to a linear regression or independent

Our model : Four different variable roles

- Modelling the pdf h :

$$\mathbf{x} \in \mathbb{R}^Q \mapsto f_{\text{clust}}(\mathbf{x}^S|K, m, \alpha)\, f_{\text{reg}}(\mathbf{x}^U|r, a+\mathbf{x}^R\beta, \Omega)\, f_{\text{indep}}(\mathbf{x}^W|\ell, \gamma, \tau)$$

  - relevant variables(S) : Gaussian mixture density

$$f_{\text{clust}}(\mathbf{x}^S|K, m, \alpha) = \sum_{k=1}^{K} p_k \Phi(\mathbf{x}^S|\mu_k, \Sigma_k)$$

  - redundant variables (U) : linear regression of $\mathbf{x}^U$ on $\mathbf{x}^R$ ($R \subseteq S$)

$$f_{\text{reg}}(\mathbf{x}^U|r, a + \mathbf{x}^R\beta, \Omega) = \Phi(\mathbf{x}^U|a + \mathbf{x}^R\beta, \Omega_{(r)})$$

  - independent variables (W) : Gaussian density

$$f_{\text{indep}}(\mathbf{x}^W|\ell, \gamma, \tau) = \Phi(\mathbf{x}^W|\gamma, \tau_{(\ell)})$$

SRUW model

- It is assumed that h can be written

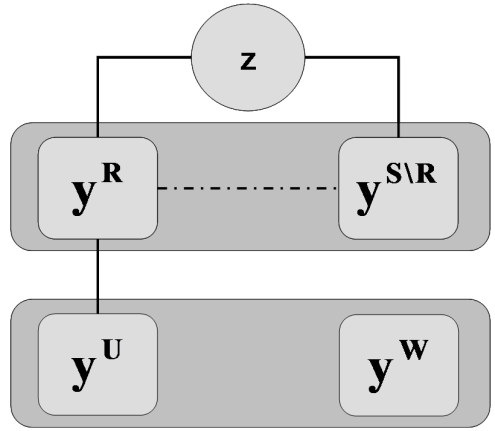
$$\mathbf{x} \in \mathbb{R}^Q \mapsto f_{\text{clust}}(\mathbf{x}^S | K, m, \alpha)\, f_{\text{reg}}(\mathbf{x}^U | r, a + \mathbf{x}^R \beta, \Omega)\, f_{\text{indep}}(\mathbf{x}^W | \ell, \gamma, \tau)$$

  - relevant variables (S) : Gaussian mixture pdf
  - redundant variables (U) : linear regression of $\mathbf{x}^U$ with respect to $\mathbf{x}^R$
  - independent variables (W) : Gaussian pdf

- Model collection :

$$\mathcal{N} = \left\{ (K, m, r, \ell, \mathbf{V});\ \begin{array}{l} (K, m) \in \mathcal{T},\ \mathbf{V} \in \mathcal{V} \\ r \in \{[LI], [LB], [LC]\}, \ell \in \{[LI], [LB]\} \end{array} \right\}$$

$$\text{where } \mathcal{V} = \left\{ \begin{array}{l} (S, R, U, W); \\ S \sqcup U \sqcup W = \{1, \ldots, Q\} \\ S \neq \emptyset, R \subseteq S \\ R = \emptyset \text{ if } U = \emptyset \text{ and } R \neq \emptyset \text{ otherwise} \end{array} \right\}$$



Model selection criterion

- Variable selection by maximising the integrated likelihood

$$(\hat{K}, \hat{m}, \hat{r}, \hat{\ell}, \hat{\mathbf{V}}) = \underset{(K,m,r,\ell,\mathbf{V})\in\mathcal{N}}{\text{argmax}} \ \text{crit}(K, m, r, \ell, \mathbf{V}) \text{ where}$$

$$\begin{aligned} \text{crit}(K, m, r, \ell, \mathbf{V}) &= \text{BIC}_{\text{clust}}(\mathbf{y}^S|K, m) + \\ &\quad \text{BIC}_{\text{reg}}(\mathbf{y}^U|r, \mathbf{y}^R) + \text{BIC}_{\text{ind}}(\mathbf{y}^W|\ell) \end{aligned}$$

- Theoretical properties :
  - The model collection is identifiable,
  - The selection criterion is consistent.

Selection algorithm (SelvarclustIndep)

- It makes use of two embedded (for-back)ward stepwise algorithms.
- 3 situations are possible for a candidate variable j :
  - M1 : $f_{\text{clust}}(\mathbf{y}^S, \mathbf{y}^j | K, m)$
  - M2 : $f_{\text{clust}}(\mathbf{y}^S | K, m)\ f_{\text{reg}}(\mathbf{y}^j | [LI], \mathbf{y}^{\widetilde{R}[j]})$ where $\widetilde{R}[j] = R[j] \subseteq S, \widetilde{R}[j] \neq \emptyset$.
  - M3 : $f_{\text{clust}}(\mathbf{y}^S | K, m)\ f_{\text{indep}}(\mathbf{y}^j | [LI])$ i.e. $f_{\text{clust}}(\mathbf{y}^S | K, m)\ f_{\text{reg}}(\mathbf{y}^j | [LI], \mathbf{y}^{\widetilde{R}[j]})$ where $\widetilde{R}[j] = \emptyset$.
- It reduces to comparing

$$f_{\text{clust}}(\mathbf{y}^S, \mathbf{y}^j | K, m) \text{ versus } f_{\text{clust}}(\mathbf{y}^S | K, m) f_{\text{reg}}(\mathbf{y}^j | [LI], \mathbf{y}^{\widetilde{R}[j]})$$

$$\Longrightarrow \textit{algorithm SelvarClust (SR model)}$$

$$\text{and} \begin{cases} j \text{ in model M2} & \text{if } \widetilde{R}[j] \neq \emptyset \\ j \text{ in model M3} & \text{otherwise} \end{cases}$$

Synopsis of the backward algorithm

1. For each mixture model (K, m) :

Step A- Backward stepwise selection for clustering :

- Initialisation : $S(K, m) = \{1, \ldots, Q\}$
- exclusion step *(remove a variable from S)*
- inclusion step *(add a variable in S)*

} using backward stepwise variable selection for regression ($\star$)

$\Rightarrow$ two-cluster partition of the variables in $\hat{S}(K, m)$ and $\hat{S}^c(K, m)$.

Step B- $\hat{S}^c(K, m)$ is partitioned in $\hat{U}(K, m)$ and $\hat{W}(K, m)$ with ($\star$)

Step C- for each regression model form r :

- selection with ($\star$) of the variables $\hat{R}(K, m, r)$
- for each independent model form ℓ : estimation of the parameters $\hat{\theta}$ and calculation of the criterion

$$\widetilde{\text{crit}}(K, m, r, \ell) = \text{crit}(K, m, r, \ell, \hat{S}(K, m), \hat{R}(K, m, r), \hat{U}(K, m), \hat{W}(K, m)).$$

2. Selection of $(\hat{K}, \hat{m}, \hat{r}, \hat{\ell})$ maximising $\widetilde{\text{crit}}(K, m, r, \ell)$

Selection of the model $\left(\hat{K}, \hat{m}, \hat{r}, \hat{\ell}, \hat{S}(\hat{K}, \hat{m}), \hat{R}(\hat{K}, \hat{m}, \hat{r}), \hat{U}(\hat{K}, \hat{m}), \hat{W}(\hat{K}, \hat{m})\right)$

Alternative sparse clustering methods

Model-based regularisation

Zhou and Pan (2009) propose to minimise a penalized log-likelihood through an EM-like algorithm with the penalty

$$p(\lambda) = \lambda_1 \sum_{k=1}^{K} \sum_{j=1}^{Q} |\mu_{jk}| + \lambda_2 \sum_{k=1}^{K} \sum_{j=1}^{Q} \sum_{j'=1}^{Q} |\Sigma^{-1}_{k;jj'}|.$$

Sparse clustering framework

Witten and Tibshirani (2010) define a general criterion $\sum_{j=1}^{Q} w_j f_j(y^j, \theta)$ with $||\mathbf{w}||^2 \leq 1, ||\mathbf{w}||_1 \leq s, w_j \geq 0 \forall j$, where f_j measures the clustering fit for variable j.

Example : for sparse K-means clustering, we have

$$f_j = \sum_{j=1}^{Q} w_j \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d^j_{ii'} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} d^j_{ii'} \right).$$

Comparing sparse clustering and MBC variable selection

Simulation	Method	CER	card($\hat{s}$).
$n = 30, \delta = 0.6$	SparseKmeans	0.40($\pm$0.03)	14.4($\pm$1.3)
	Kmeans	**0.39**($\pm$**0.04**)	25.0($\pm$0)
	SU-LI	0.62($\pm$0.06)	22.2($\pm$1.2)
	SRUW-LI	0.40($\pm$0.03)	8.1($\pm$1.9)
$n = 30, \delta = 1.7$	SparseKmeans	**0.08**($\pm$**0.02**)	8.2($\pm$0.8)
	Kmeans	0.25($\pm$0.01)	25.0($\pm$0)
	SU-LI	0.57($\pm$0.03)	23.1($\pm$0.2)
	SRUW-LI	**0.085**($\pm$**0.08**)	6.8($\pm$1.4)
$n = 300, \delta = 0.6$	SparseKmeans	0.38($\pm$0.003)	24.00($\pm$0.5)
	Kmeans	0.36($\pm$0.003)	25.0($\pm$0)
	SU-LI	0.37($\pm$0.03)	25.0($\pm$0)
	SRUW-LI	**0.34**($\pm$**0.02**)	7.0($\pm$1.7)
$n = 300, \delta = 1.7$	SparseKmeans	**0.05**($\pm$**0.01**)	25.0($\pm$0)
	Kmeans	0.16($\pm$0.06)	25.0($\pm$0)
	SU-LI	0.05($\pm$0.01)	14.6($\pm$2.0)
	SRUW-LI	**0.05**($\pm$**0.01**)	5.6($\pm$0.9)

Results from 20 simulations with $Q = 25$ and card(s) $= 5$

Comparing sparse clustering and MBC variable selection

Fifty independent simulated data sets with $n = 2000$, $Q = 14$, the first two variables are a mixture of 4 equiprobable spherical Gaussian : $\mu_1 = (0, 0)$, $\mu_2 = (4, 0)$, $\mu_3 = (0, 2)$ and $\mu_4 = (4, 2)$.
$\mathbf{y}_i^{\{3,\ldots,14\}} = \tilde{a} + \mathbf{y}_i^{\{1,2\}}\tilde{\beta} + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \tilde{\Omega})$ and $\tilde{a} = (0, 0, 0.4, \ldots, 4)$ and 2 different scenarios for $\tilde{\beta}$ and $\tilde{\Omega}$.

Method	Scenario 1	Scenario 2
Sparse Kmeans	0.47 ($\pm$ 0.016)	0.31 ($\pm$ 0.035)
Kmeans	0.52 ($\pm$ 0.014)	0.57 ($\pm$ 0.015)
SR-LI	0.39 ($\pm$ 0.039)	0.42 ($\pm$ 0.082)
SRUW-LI	**0.57 ($\pm$ 0.04)**	**0.60 ($\pm$ 0.015)**

The adjusted Rand index

Method	Scenario 1	Scenario 2
Sparse Kmeans	14 ($\pm$ 0)	13.5 ($\pm$ 1.5)
Kmeans	14 ($\pm$ 0)	14 ($\pm$ 0)
SU-LI	12 ($\pm$ 0)	3.96 ($\pm$ 0.57)
SRUW-LI	2 ($\pm$ 0.20)	2 ($\pm$ 0)

The number of selected variables

Variable selection in a supervised Classification context

We turn now to an other variable selection problem.

- Aim : classify observations described with Q variables in one of K groups given a priori
- The classifier is designed from a training sample

$$\{(\mathbf{y}_1, z_1), \ldots, (\mathbf{y}_n, z_n); \mathbf{y}_i \in \mathbb{R}^Q, z_i \in \{1, \ldots, K\}\}$$

  where the labels $z_i, i = 1, \ldots, n$ are known.
- We consider here generative models which assume a parameterised form for the group conditional density $f(\mathbf{y}_i | z_i = k)$.
- From which, it follows that the density of the $\mathbf{y}_i$ is a mixture density with K components.
- In such a decision-making context, variable selection is often crucial to design an efficient classifier.

Variable selection for Gaussian Classifiers

- The classifier is designed from a training sample

$$\{(\mathbf{y}_1, z_1), \ldots, (\mathbf{y}_n, z_n); \mathbf{y}_i \in \mathbb{R}^Q, z_i \in \{1, \ldots, K\}\}$$

- Gaussian generative model :

$$\begin{cases} f(\mathbf{y}_i | z_i = k, m) = \Phi(\mathbf{y}_i | \mu_k, \Sigma_k), \ \forall i \in \{1, \ldots, n\} \\ P(z_i = k) = p_k \end{cases}$$

  - LDA : $m = [LC]$ $(\forall k, \Sigma_k = \Sigma)$
  - QDA : $m = [L_k C_k]$
  - EDDA 14 models derived from the eigenvalue decomposition of the group variance matrices.

- Variable selection can be proceeded with the SRUW model in a simple way since the classification is known.
- The resulting (for-back)ward procedures generalise the standard variable selection procedures for LDA. (Murphy *et al. 2010*, Maugis *et al. 2010*)

Illustrations of variable selection in a supervised setting

Landsat Satellite Data set

It consists of the multi-spectral values of pixels in a tiny sub-area of a satellite image. The data points are in $\mathbb{R}^{26}$ and split into six classes. The original learning set has 4435 samples and a test set with 2000 samples is available.

- LDA and QDA are compared.
- 1000 samples randomly selected 100 times from the training data are used to estimate and select the model.
- The same 12 variables are selected for both models in average ; $\hat{R} = \hat{S}$ ($\hat{r} = [LC]$), and $\hat{W} = \emptyset$.

with variable selection		without variable selection	
LDA	QDA	LDA	QDA
21.00 $\pm$ 0.53	16.21 $\pm$ 0.68	18.05 $\pm$ 0.48	17.90 $\pm$ 0.57

Averaged classification error rate

Illustrations of variable selection in a supervised setting

Leukemia data set

These data come from a study of gene expression divided in two types of acute leukemias : 47 tumor samples for acute lymphoblastic leukemia (ALL) and 25 for acute myeloid leukemia (AML) measured on $Q = 3571$ genes.

We analyze the Leukemia data set using 38 (27 are ALL and 11 are AML) samples in the training set and 34 (20 are ALL and 14 are AML) samples in the test set.

Models	LDA	QDA	$[L_k C]$
card($\hat{S}$)	8	8	3
card($\hat{R}$)	2	2	3
card($\hat{U}$)	3058	2848	1912
card($\hat{W}$)	505	715	1656
Misc. test obs. (ALL, AML)	(2,4)	(0,0)	(0,0)

Variable selection and misclassification error rate.

Discussion

Interest of variable selection

- In the unsupervised setting, variable selection is essentially useful to interpret the clustering.
- In the supervised setting, variable selection could improve dramatically the perfomances of quadratic classifiers.

Backward or Forward selection ?

- Backward selection can be expected to provide more stable results.
- Forward selection is necessary in high dimension settings.

Softwares

Free softwares can be downloaded from the Cathy Maugis home page

`http ://www.math.univ-toulouse.fr/~maugis`


Celeux, C., Martin-Magniette, M.-L., Maugis, C., and Raftery, A. E. (2011).
Letter to the editor in relation with a framework for feature selection in clustering.
Journal of the American Statistical Association, 106.


Law, M. H., Figueiredo, M. A. T., and Jain, A. K. (2004).
Simultaneous feature selection and clustering using mixture models.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(9) :1154–1166.


Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009a).
Variable Selection for Clustering with Gaussian Mixture Models.
Biometrics, 53(3872) :3882.


Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009b).
Variable selection in model-based clustering : A general variable role modeling.
Computational Statistics and Data Analysis, 65(701) :709.


Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2011).
Variable selection in model-based discriminant analysis.
Journal of Multivariate Analysis.
in revision.


Murphy, T. B., Dean, N., and Raftery, A. (2010).
Variable Selection and Updating In Model-Based Discriminant Analysis for High-Dimensional Data.
Annals of Applied Statistics, (4) :396–421.


Raftery, A. E. and Dean, N. (2006).
Variable Selection for Model-Based Clustering.
Journal of the American Statistical Association, 101(473) :168–178.


Witten, D. M. and Tibshirani, R. (2010).
A framework for feature selection in clustering.
Journal of the American Statistical Association, 105(490) :713–726.


Zhou, H. and Pan, W. (2009).
Penalized model-based clustering with unconstrained covariance matrices.
Electronic Journal of Statistics, 3 :1473–1496.