Discrimination through the Regularized Nearest Cluster Method. (1992) (Invited Paper at the 10th Symposium on Computational Statistics [COMPSTAT], Neuchâtel, Switzerland), in: *Computational Statistics,* Y.Dodge and J. Whittaker,eds., Physica Verlag, Vienna p 103-118.

Discrimination through the regularized nearest cluster method.

Ludovic Lebart

Centre National de la Recherche Scientifique
E.N.S.T., 46 Rue Barrault, 75013, Paris, France.

Abstract

This paper contains three parts. The first part consists of a brief review of the discrimination techniques used when dealing with large arrays of sparse qualitative data. The second part presents the "Regularized Nearest Cluster Method", an efficient and versatile technique of discrimination, well adapted to this kind of data. This technique is compared to some other existing methods likely to be used in similar contexts. The third part briefly discusses the interest of these methods in the domain of textual data analysis.

1. BASIC CONCEPTS AND NOTATIONS

Let p designate the number of variables, n the number of individuals, and K the number of categories to which belong these individuals.

The most classical discrimination techniques are based on the normal distribution, whose density function is written, μ_k and Σ_k designating the theoretical mean vector and covariance matrix of class k :

$$f_k(\mathbf{X}) = (2\pi)^{-p/2} \left|\Sigma_k\right|^{-1/2} \exp\{-1/2(\mathbf{X}-\mu_k)'\Sigma_k^{-1}(\mathbf{X}-\mu_k)\}$$

Replacing the theoretical values by their estimates, one can derive the discriminant scores and the Mahalanobis distances :

$d_k(\mathbf{X}) = (\mathbf{X}-\mathbf{m}_k)' \mathbf{S}_k^{-1}(\mathbf{X}-\mathbf{m}_k) + \ln\left|\mathbf{S}_k\right|$ (discriminant score for group k, when the prior probabilities of the groups are equal)

$d_k(\mathbf{X}) = (\mathbf{X}-\mathbf{m}_k)' \mathbf{S}^{-1}(\mathbf{X}-\mathbf{m}_k)$ (global Mahalanobis distance, directly related with linear discriminant analysis)

$d_k(\mathbf{X}) = (\mathbf{X}-\mathbf{m}_k)' \mathbf{S}_k^{-1}(\mathbf{X}-\mathbf{m}_k)$ (local Mahalanobis distance, corresponding to quadratic discriminant anaysis)

with $\mathbf{S_k} = (1/n_k) \sum_{i=1}^{n_k} (\mathbf{x}_i - \mathbf{m}_k)'(\mathbf{x}_i - \mathbf{m}_k)$, for $\mathbf{x_i} \in$ category **k,** containing n_k individuals.

$\mathbf{m}_k = (1/n_k) \sum \mathbf{x_i}$, for $\mathbf{x_i} \in$ category k.

$\mathbf{m} = (1/n) \sum \mathbf{x_i}$, for all $\mathbf{x_i}$, with $n = \sum_k n_k$

In the regularised discriminant method as proposed by J.Friedman [7], a new estimate $\mathbf{S}_k(\lambda,\gamma)$ is computed for each local covariance matrix, rendering it similar to the global covariance matrix (role of weight λ), and also to a multiple of the identity matrix (role of weight γ) :

$$\mathbf{S}_k(\lambda,\gamma) = (1 - \gamma)\, \mathbf{S}_k(\lambda) + (\gamma/p)\ \mathrm{tr}\,[\mathbf{S}_k(\lambda)]\, \mathbf{I}$$

$$\text{with } \mathbf{S}_{k(\lambda)} = \frac{(1-\lambda)\mathbf{S}_k + \lambda\mathbf{S}}{(1-\lambda)n_k + \lambda n}$$

See also [6] for a different technique of estimation of the parameters λ and γ.

These techniques provide interesting results in the case of small or medium-sized data sets, when the initial problem is ill-posed ($n \leq p$) or poorly posed ($n > p$, but still comparable to p).

In the case of large sparse matrices, however, the scale of the phenomenon creates new problems : there is a need for understanding what happen in the high-dimensional spaces...

Is it worthwhile keeping all the principal axes ? Is there some kind of filtering adapted to high-noise level data ?

2. REGULARIZATION THROUGH DISCARDING THE LAST PRINCIPAL AXES

From a numerical point of view, diagonalizing is a much safer operation than computing inverse matrices. Perturbation theory proves that the stability of eigenvectors is a function of the differences between consecutive eigenvalues. In this context, it could be advisable to discard the dimensions corresponding to the smallest eigenvalues, which are very sensitive to minor changes in the input data set [3] .

The purpose of sections 2 and 3 is to contribute to discriminant analysis in this setting.

2.1 Principal Axes of the whole sample

The reduction technique occurring during the first step depends on the nature and the statistical properties of the input data. In the example presented throughout this paper, this first phase consists in a Correspondence Analysis (C.A.), adapted to the sparse contingency tables under study (Note that a regular Singular Value Decomposition could be used as well).

The new coordinates of individual i on the principal axis r of the whole sample eigen-analysis, are designated by y_{ri} , .

$$y_{ri} = \mathbf{u'}_r (\mathbf{x}_i - \mathbf{m}),$$

where $\mathbf{u}_r$ is the r^{th} normalized eigenvector of **S** corresponding to the eigenvalue α_r, and the r^{th} column of the (p,R) matrix **U**, (R is the number of selected eigenvalues).

The usual squared euclidean distance of any point i to the centroid k (including those points not belonging to category k and those not belonging to the learning sample) is written as :

$$d^2_k(i) = \sum_j (x_{ij} - m_{kj})^2). \qquad (1)$$

This distance is written for the new base :

$$d^2_k(i) = \sum_r (y_{ri} - \bar{y}_{kr})^2), \quad \text{with } \bar{y}_{kr} = \mathbf{u'}_r (\mathbf{m}_k - \mathbf{m}) \qquad (2)$$

whereas the global Mahalanobis squared distance (linear discriminant analysis) could be written as :

$$D^2_k(i) = \sum_r (y_{ri} - \bar{y}_{kr})^2/\alpha_r), \qquad (3)$$

$D^2_k(i)$ is regularized if $R \leq p' \leq \text{Min } (n,p)$ (p' designates the rank of the data matrix)

2.2 Principal axes of sub-groups

For each sub-group k, the (R,R) covariance matrices are computed separately, the input data being the coordinates computed previously.

The new coordinates of an individual on the principal axis s of the eigen-analysis (a Principal Component Analysis) performed within the group k are :

$$z_{ski} = \mathbf{v'}_{sk} (\mathbf{y}_i - \bar{\mathbf{y}}_k),$$

where $\mathbf{v}_{sk}$ is the s^{th} normalized eigenvector of $\mathbf{U'S_kU}$ corresponding to the eigenvalue β_{sk}.

This formula of projection onto axis s holds also for the points not belonging to category k (*supplementary* or *illustrative* elements, according to the terminology of descriptive multivariate analyses).

One can find again the same usual distances computed in each of these K new bases, for any point i, to the centroid k (including those not belonging to category k and those not belonging to the learning sample).

$$d^2_k(i) = \sum_S (z_{ski} - \bar{z}_{sk})^2, \qquad (4)$$

(S being the number of selected axes at this step).

whereas the local Mahalanobis distance (quadratic discriminant analysis) could be written :

$$D^2_k(i) = \sum_S (z_{ski} - \bar{z}_{sk})^2/\beta_{sk}\,, \qquad (5)$$

Such distance is "regularised", if $R \leq p' \leq Min(n,p)$ (p' designates the rank of the data matrix) and could be said to be "doubly regularised" if $S < R$, (see remark below).

Note that if $S = R = p$, the distances given by formulas (1), (2),and by the K formulas (4) (there are in fact K different orthonormal bases, therefore K different formulas) are equal.

The empirical study which follows will focus on the effects of the dimensions of these subspaces on the misclassification rates, in both learning and test samples.

Remark:

There are in fact two ways of computing regularized Mahalanobis distances by truncation of the p-dimensional space :

1- To reduce the global dimensions of the global space by keeping the R first eigenvectors, then to re-compute for each value of R the K classes covariance matrices.

K diagonalizations (or inversions, if possible) are then necessary at each step. This is the solution adopted throughout this paper.

2 - To compute (and diagonalize) each class covariance matrix $\mathbf{S}_k$ in the full initial p-dimensional space, and discard the last axes for each group separately. This option does not imply a re-computation of the eigenvalues at each step. However, these new Mahalanobis distances are not computed in a common space, and cannot be compared as in the usual framework of quadratic discriminant analysis.

2.3 The example data set

The data set that will serve as example throughout the various sections of this paper consists of a (634 x 83) sparse binary table containing 4039 non-zero elements (4039 occurrences of p =83 words used in n = 634 responses to an open-ended question).

The set of the 634 rows (respondents) is broken down into three categories of age. The purpose is to classify these categories from the words used in answering the open question. The real scope of the problem is presented in section 4.

Our current criterion of goodness of discrimination is the percentage of successes, which will be systematically computed for both test samples and learning samples. Throughout this paper, the test sample consists of one third (211 individuals) of the full sample drawn at

random from the whole sample.

The first step is a change of the axes performed through C.A. Special algorithms adapted to the sparsity of the matrix have been developed [10] and could be used.

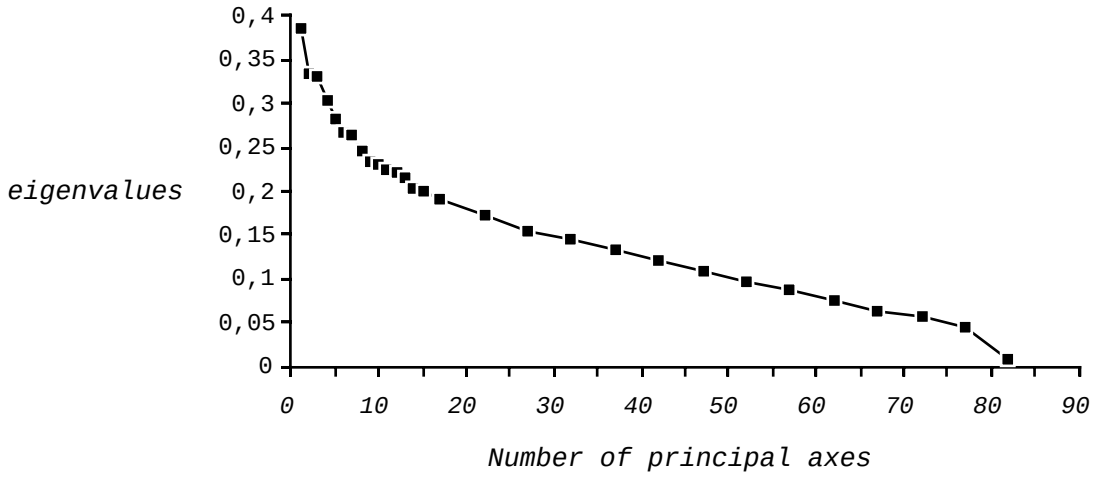


Figure 1. Eigenvalues of the (83,83) sparse matrix

The sequence of eigenvalues shown in Figure 1 is typical of a sparse binary matrix : the decrease of the values is extremely slow, almost linear after axis number 15. The first 15 eigenvalues account for 37% of the trace. Each of the remaining axes corresponds approximately to 1% of the trace.

Figure 2 shows the "trajectories" of the percentages of successes obtained for each of the three previous distances : Usual euclidean distance (square symbols), Global Mahalanobis distance (triangular symbols), Local Mahalanobis distance (diamond shaped symbols).

We note that the rates corresponding to learning samples success rates (black symbols) increase continuously with the number of axes, whereas those rates corresponding to the test samples are almost stabilized beyond 15 axes.

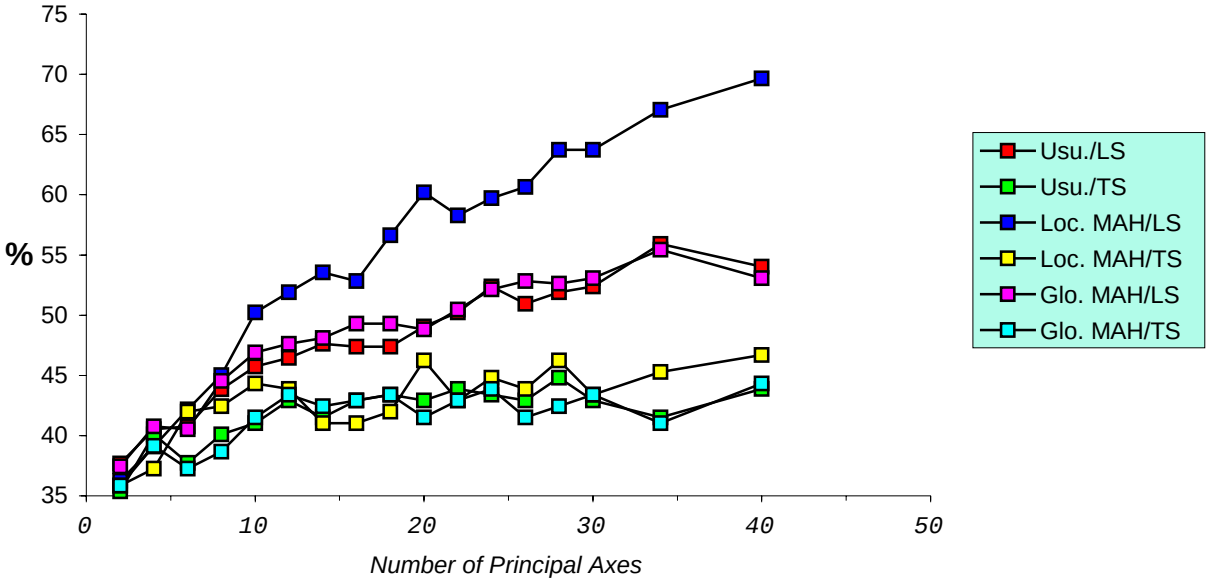
Among the learning sample trajectories, the local Mahal. distance increases vigorously, and reaches a level of 70% of successes for 40 axes. Such distance depending upon an increasing number of parameters obviously "sticks" to the real data, without comparable improvement for the test sample.

The global Mahal. distance is very similar to the usal euclidean distance, although leading to slightly better percentages of successes. The situation is far less clear for the test samples. One can guess however an average superiority of the local Mahal. distance.

Figure 3 is a zoom of the lower part of figure 2, dealing with the three test samples results. If we except two isolated points and the interval ranging from axes 14 to 18, the dominance

of the local Mahalanobis distance is clear.

Figure 2. Learning and Test Sample percentages of successes for Usual, Global Mahal. and Local Mahal. distances



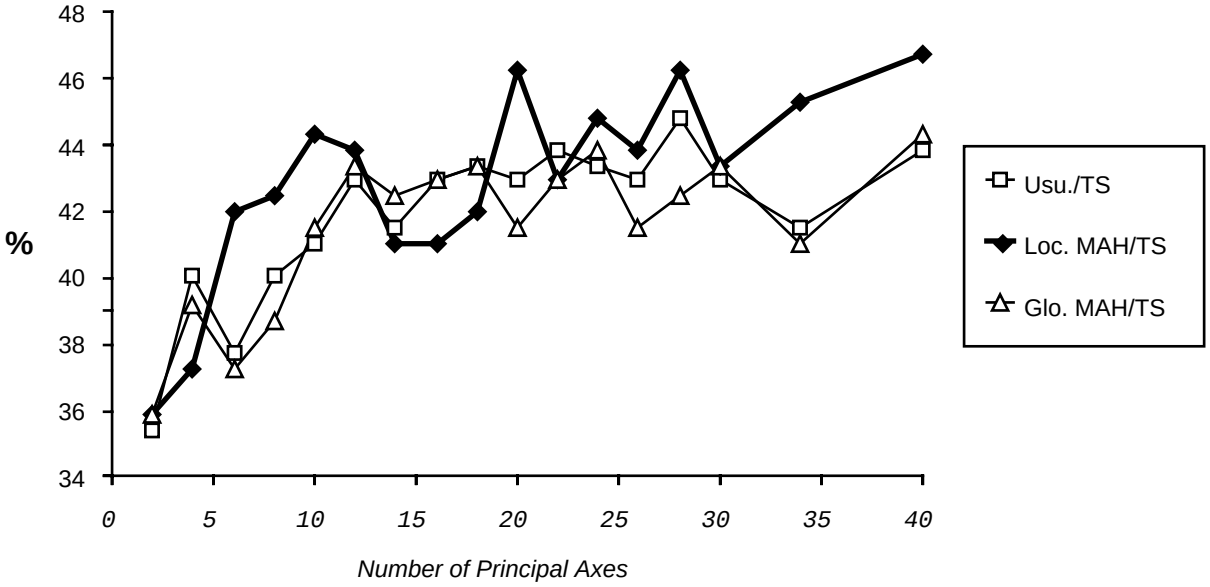


Figure 3. Percentages of successes (Test sample only) (usual, local Mahal. and Global Mahal. distances)

The fluctuations, however, are surprisingly large. We must keep in mind that K

diagonalizations occur for each value of R. It may happen that for some values of R, the instability is created by the smallest eigenvalues β_{sk} (see formula 5) of the local covariance matrix of group k computed in the new reduced base.

3. THE REGULARIZED NEAREST CLUSTER METHOD

This discrimination technique takes advantage of the shape of the swarm of n points in the original p-dimensional space. The weakness of the previous assignment rules lies in the computation of the distances of points to a unique point for each category, whatever the sophistication of the used distance. The new proposed assignment rule will take into account the shape of the "sub-cloud" corresponding to each category.

3.1 Basic principles

The regularized nearest cluster method proceeds in four steps :

1) A priori clustering of the whole sample according to the appropriate distances computed in the original space (regularization by truncation using principal axes will be dealt with later on).

For this kind of large sparse data, a procedure of "mixed or hybrid clustering" is recommended [12]. Let us recall that this technique comprises 4 phases : a) Reduction of the number of individuals through a k-means technique. b) Hierarchical classification (using Ward's criterion) of the provisional groups obtained. c) Choice of the cutting level of the dendrogram. d) Iterative reassignment of the individuals to optimise locally the partition defined by the previous cut.

As a result, a "basic partition" with q classes is obtained.

2) A new provisional target variable with qK categories is obtained, *for the learning sample only*, through cross-tabulation of two nominal variables : the basic partition obtained during the first step on the one hand, the initial target variable on the other. The kp centroids of these classes are computed and stored.

3) The individuals or objects of both learning-sample and test-sample are then assigned to the nearest centroid. In the following examples, the usual euclidean distance is used.

4) Individuals are eventually assigned to the K initial categories to which belongs the cluster whose centroid has been selected (since the qK provisional groups are included in these K categories).

In practice, the first step directly produces a series of partitions, corresponding to different cuts of the dendrogram.

Two parameters play an important role in contributing to the quality of the discrimination :

- The number of classes of the basic partition obtained during the first step,

- The number of principal axes kept to compute the distances.

The choice of the final model depends on these two parameters whose values are customized to real situations by maximizing a sample based estimate of the rate of successes.

Two series of empirical results tend to prove that the task of determination of these two parameters involves a moderate amount of computation.

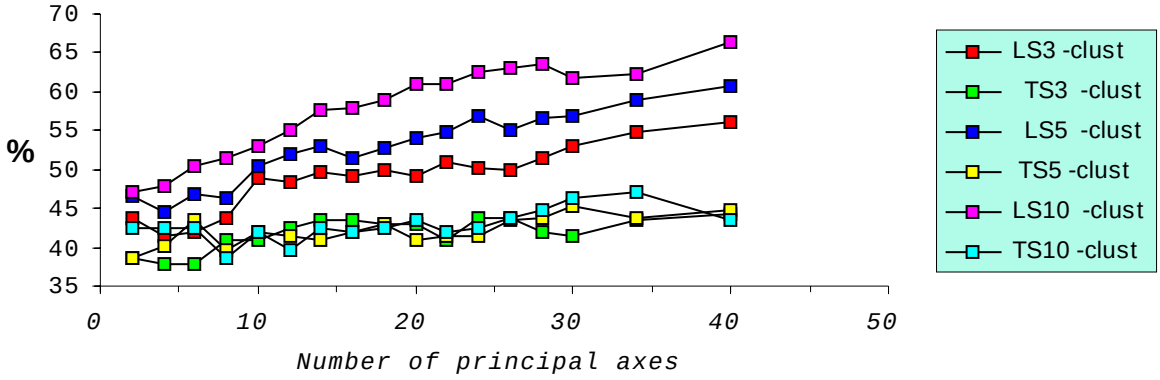
1) There is a relatively early stabilization of the rate of successes (for the test-sample) when the number of axes increases (between 5 and 20 axes, out of 100) for most of our examples.

2) The number of classes of the basic partition corresponding to the highest rates of successes remains small (less than 10 in general, 4 for our example).

Although these results need to be assessed by mathematical proof, the order of magnitude should prompt more experiments on real data as well as simulated data.

Figure 4 allows one to check a predictable result : The larger the number of classes of the basic partition, the larger the percentages of successes *on the learning sample*.

Figure 4. Percentages of successes. Regularized Nearest Cluster method. Partitions with 3,5,10 classes.



An increase of the number of classes of the basic partition enhances the quality of the fit of the learning sample. This "forced" fit is however unable to seize new structural features that would improve the predictive power of the method. Once again, the trajectories of the test samples success rates are far from following the same trend.

If we compare, on the same test-sample, the percentages of successes obtained from the best method of section 2 (discrimination using a local Mahal. distance) to some of those

percentages issued from the regularized nearest cluster method, we observe similar performances (figure 5).

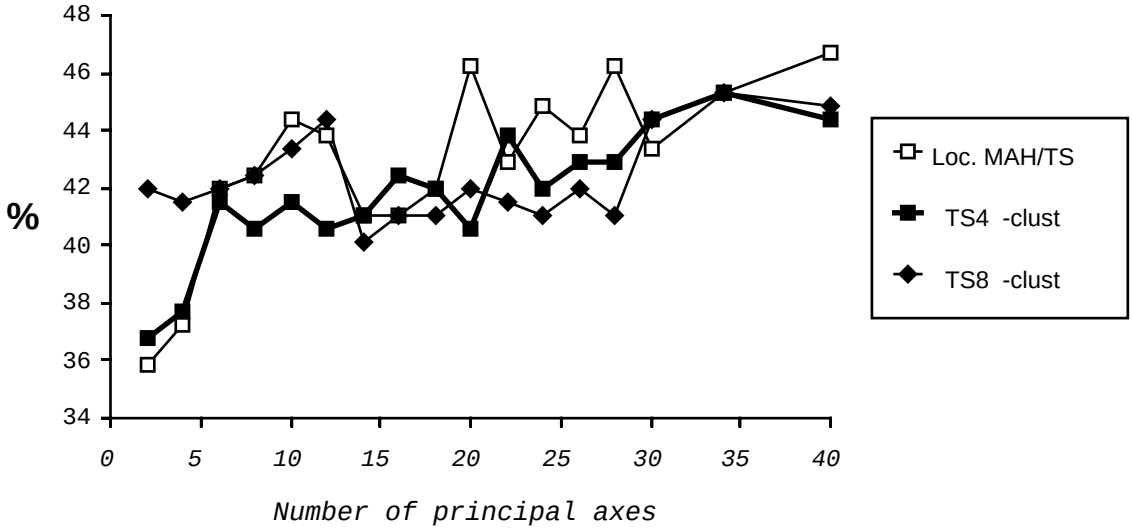


Figure 5. Percentages of successes on the same test sample (Local Mahal. distance compared to RNCM with 4 and 8 classes)

Up to now, RNCM is carried out using the usual enclidean distance. The good results obtained by the local Mahal. distance within the range of the axes 20 - 30 are an incentive to experiment with a variant of the proposed NCRM method making use of this distance. However, this distance would increase considerably the amount of computation required to optimize the parameters q and R (see below).

3.2 Determination of the two parameters q and R

The sample-based estimates of the parameters p and R are determined as follows :

We start from T partitions (partition t has q_t classes). This sequence of partitions is not nested, since each partition is locally optimized by reallocation of the individuals after the cut of the dendrogram.

For each of the m learning samples (m = 30 in the following numerical application), the distances of each of the n individuals to the Sqt centroids are computed.

The whole operation is done for each dimensionality of the subspace : 1,2,3,...R.

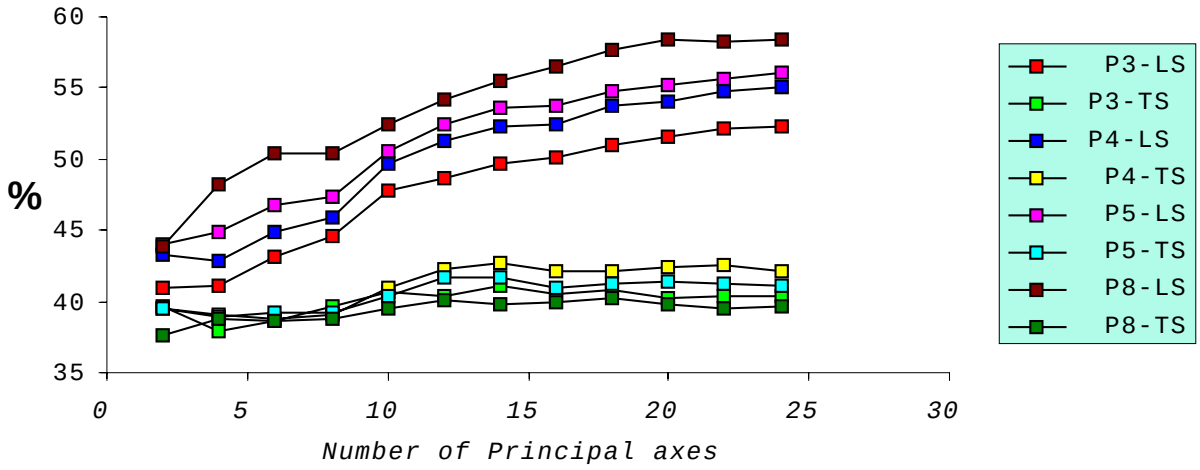
Figure 6 shows the average percentages of successes in the case of m = 30 (30 drawings at random of test-samples and learning samples).

The maximum dimension is R = 25 (all the nested subspaces spanned by the 2,3,...25, first principal axes are considered).

All the partitions with 2, 3,..,8 classes are taken into account, though only a selection of them has been represented on the graphical display.

It is satisfactory to note the effective stabilization of the trajectories of both learning samples and test samples.

Figure 6. Average trajectories of successes percentages for 4 sizes of the basic partitions (3,4,5,8 classes)



The most interesting results concerns the fact that the number of classes leading to the best efficiency of the test sample (q = 4) is distinct from its analogue of the learning sample (q = 8).

For the learning sample, as stated above, the success rate is automatically related to the number of classes of the basic partitions. As expected, the test sample behaviour is more subtle. Such observation suggests that some important structural features have been taken into account by the 4-classes partition . To split or to merge these classes has a negative influence on the discrimination power of the method. Such result could be of the utmost interest for the user, since it may help to comprehend the mechanism of the discrimination rule.

Figure 7 presents a zoom of the success rates corresponding to the test samples only, while table 1 contains an excerpt of the numerical results serving as a basis for the selection process of the parameters.

The chosen parameters are thus q = 4 and R = 14. This choice is corroborated by the general aspects of the curves as they can be observed on figure 7 (regular trends, clear-cut separations).

The corresponding rate of successes has the value 42.64. The estimated standard deviation

of this value is 0.48 ; therefore, we have the approximate confidence interval [41.68 , 43.59] (p = 0.05), for the rate of successes computed within a test-sample made of 33% of the entire set of individuals.

This system of test-samples leads obviously to pessimistic estimate of the risk of misclassification, since it notably reduces the size of the learning sample. A series of results based on the leave-one-out technique will complement these figures.

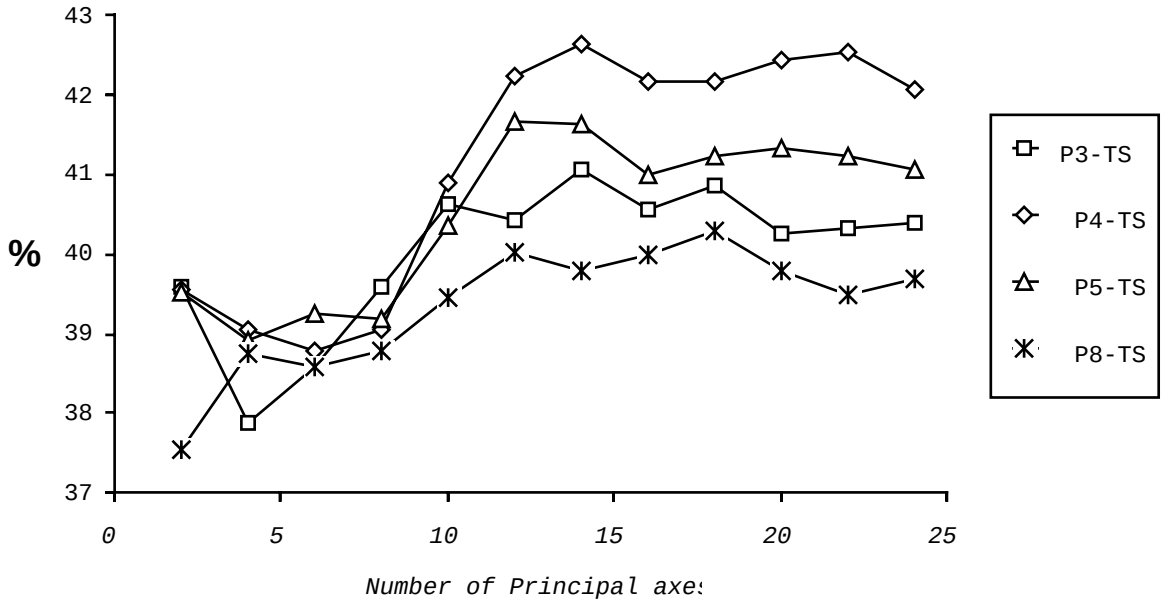


Figure 7. Average trajectories (30 test samples) of successes rates for RNCM with 4 sizes of basic partitions

TABLE 1. AVERAGE SUCCESS RATES FOR LEARNING AND TEST-SAMPLES

	PART 3		PART 4		PART 5		PART 6		PART 8	
AXES	LS	TS	LS	TS	LS	TS	LS	TS	LS	TS
2	41.00	39.60	43.31	39.54	44.08	39.51	44.29	39.17	43.88	37.55
4	41.15	37.87	42.90	39.04	44.88	38.91	46.47	39.79	48.29	38.76
6	43.20	38.57	44.83	38.79	46.76	39.27	47.85	38.36	50.36	38.58
8	44.56	39.60	45.93	39.05	47.41	39.17	48.41	38.11	50.40	38.79
10	47.78	40.63	49.63	40.90	50.51	40.35	51.24	39.41	52.46	39.44
12	48.66	40.43	51.27	42.23	52.49	41.68	53.04	40.72	54.17	40.03
14	49.64	41.08	52.36	42.65	53.65	41.63	53.98	40.25	55.44	39.80
16	50.12	40.56	52.42	42.16	53.74	41.01	53.78	39.48	56.49	39.98
18	51.05	40.88	53.78	42.16	54.71	41.23	55.27	40.50	57.66	40.29
20	51.56	40.27	54.01	42.44	55.23	41.34	55.54	40.46	58.37	39.78
22	52.16	40.34	54.83	42.55	55.64	41.23	56.42	40.36	58.29	39.50
24	52.35	40.40	55.03	42.07	56.05	41.08	57.10	40.79	58.44	39.68

Remark 1: A substantial amount of computation can be saved during the determination of tables such as Table 1 by storing the array of distances (distances between the n individuals

and the series of centroids relating to all the basic partitions) and by updating it as the number of dimensions increases. It is not necessary to re-compute the whole distances at each step. Note for instance that the computation of such table (relating to 30 test-samples) is much less costly than the computation needed to obtain figure 2 (which relates to a unique test-sample).

Remark 2: A variant of the technique proposed here consists of clustering separately the individuals within each category of the learning sample. This procedure has the advantage of describing more carefully the density of each sub-group, using for instance a variable number of clusters (and consequently a variable number of centroids) to summarize each category. In such case, the number of involved parameters increases, and the sample-based estimation of these parameters implies much more computation time, since the clustering of the individuals must be performed again for each drawing of the test sample.

4. DISCRIMINATING FROM TEXTUAL DATA (The examples data set)

This section aims at presenting the real setting of the numerical example used in the previous sections, emphasizing the importance of the processing of large sparse matrices in this domain of application.

Discriminant analysis is needed in various domains involving textual data. The most classical applications probably concern the problems related to authorship attribution. In this context, discrimination is often considered a major objective of quantitative literary works (see for instance [9]). Discriminant Analysis is often used to allocate texts (books, scientific papers, abstracts, in the framework of Automatic Information Retrieval applied to documentary databases) or individuals described by texts (Medicine, History, Marketing, see also [4]).

During the processing of sample surveys data, the technique of discriminant analysis applied to responses to open questions currently help to the post-coding of these responses (i.e. replacing one open question by one or several closed questions), to predict non-responses, to assess classifications. Our example pertains to this last category of application.

4.1 Some specific features of textual data.

The chosen basic statistical unit is the *graphical form* defined as a series of non-delimiting characters (blanks, periods, commas...). A single word can generate several graphical forms, depending on its case or its gender in the text; a single graphical form can also refer to several words.

One can also define larger units composed of several consecutive graphical forms. These units are called *repeated segments* [13], [2]. These are sequences of simple forms that appear with a frequency greater than a given threshold. Special computational algorithms are able to uncover such segments. Evidently, introducing such new units considerably increases the number of variables as well as the sparsity of the data matrix.

As usual when dealing with sparse data, two matrices **R** and **X** are involved. For a response or an individual i, row i of the matrix **R** contains the addresses r(i,j) (relative to a *vocabulary*) of the graphical forms (or segments) that constitute the response, while respecting

the order and the possible repetitions of these forms. From **R** it is therefore possible to reconstitute the original responses integrally.

Matrix **X** has the same number n of rows as **R** but has as many columns as there are graphical forms used by the entire set of individuals, that is p columns. At the intersection of row i and column j of **X** is the number of times x_{ij} graphical form j was used by individual i in his or her response. **X** can easily be constructed from **R** but the converse is not true: the information relative to the order of the forms in each response is lost in **X**.

In order to take advantage of the sparsity of **X,** the statistical and algorithmic computations that involve **X** are actually programmed with **R** [10]**.** The first global diagonalization mentionned in section 2.1 makes use of such special algorithms.

Discrimination is usually performed after various selections or groupings :

- Selection of words using a threshold of frequency : a low threshold will increase once again the number of variables and the sparsity of the matrix, but rare words could nevertheless be very productive in a discrimination process.
- Selection of words and segments according to their *specificities* (only words characteristic of certain categories are selected) [11].
- Discrimination after grouping of words or segments (see for instance [5], following the recommendation of [8]).

Most of these procedures aim at reducing the size of the input data matrix, or at solving ill or poorly posed problems. This prior reduction of the data (involving obvious losses of information) could be avoided by using the tools presented in this paper.

4.2 The context of the example

The data set serving as example is extracted from an international sample survey concerning "dietary culture" [1] . The questionnaire, containing numerous closed questions, comprises also the following open-ended question :

"*What dishes do you like and eat often?* (With a probe: "*Any other dishes you like and eat often?*")

This question has been asked in the three cities of Tokyo, New-York and Paris. The responses have been given in three different languages : Japanese, English and French.

The sub-sample relating to the city of New-York contains 634 individuals. The global corpus of open responses contains 6511 occurrences of 638 distincts words The data processing presented here takes into account the 83 words appearing at least 12 times, corresponding to 4039 occurrences. This threshold of 12 allowed us to deal with a reasonably sized matrix X (634,83) for a methodological study. The statements of section 4.1 prompt us to select a much lower threshold.

In the three countries under survey, a general typology of the respondents described by their lexical profiles highlights the prominent role of the two variables *age* and *sex*, and also of the interaction between these two basic variables. But a direct comparison by merging the different files is not possible, since the languages (and consequently the basic variables) are different.

It has been the task of these techniques of discrimination adapted to large sparse matrices to assess, for each country, the predictive power of the raw responses over various socio-economic characteristics of the individuals. Discrimination is often used in this context to validate the patterns discovered through multivariate descriptive techniques.

5. CONCLUDING REMARKS

The techniques of regularization dealt with in this paper can tackle the worst situations from a numerical point of view, since they start iteratively from the principal axes corresponding to the largest eigenvalues issued from a singular value decomposition of the input data matrix. Such data matrix must be at least of rank one ! This is the weakest theoretically possible condition, although of no practical use...

For the understanding of the phenomenon underlying the data set, it seems furthermore valuable to obtain simultaneously a good discrimination (with a risk of misclassification comparable to the risk provided by the best available blind procedures) as well as a model or a relatively parcimonious description of the data : number of dimensions, number of clusters, both of these parameters being determined to ensure the best prediction in the setting of a sample-based optimization.

In the low dimensional space obtained, marked out by a small number of clusters, it could perhaps be easier to understand the reasons why the discrimination process is working.

Aknowledgements

The author wish to thank Professor H. Akuto, (University of Tokyo), Pr T.Okamoto (Institute of Urban Life, Tokyo Gas Co.) for their permission to use some of the data from the International Survey about Dietary Culture in Tokyo, New-York, and Paris; and Dr C. Callant for carefully reading a first version of the manuscript.

6. REFERENCES

[1] Akuto H.(Ed.) (1992). *International Comparison of Dietary Cultures* (To be published), Tokyo.

[2] Bécue M. (1988). Characteristic Repeated Segments and Chains in Textual Data Analysis'. *COMPSTAT, 8th Symposium on Computational Statistics* , Physica Verlag

[3] Benzecri J.P.(1977). Analyse Discriminante et Analyse Factorielle, *Les Cahiers de l'Analyse des Données*, II,n°4,369-406.

[4] Benzécri J-P.& coll. (1981). *Linguistique et Lexicologie*, in *Pratique de l'Analyse des Données,* Tome 3, Dunod, Paris.

[5] Celeux G., Hébrail G., Mkhadri A., Suchard M. (1991). Reduction of a Large Scale and ill-conditioned Problem statistical Problem on textual Data, in *Applied Stochastic Models and Data Analysis, Proceedings of the 5th Symposium in ASMDA*, Gutierrez R. and Valderrama M.J. Eds,World Scientific, 129-137.

[6] Callant C.M. (1991). *Technique de Lissage et de Régularisation en Analyse Discriminante*, Dissertation, *Université Paris IX, Dauphine, (Publ. INRIA TU177),* Paris.
[7] Friedman J.H. (1989), Regularized Discriminant Analysis, *Journal of the American Statistical Association*, 84, 165-175.
[8] Hand D.J. (1981 and 1986). *Discrimination and Classification*, Wiley, New-York.
[9] Holmes D.I. (1985). The Analysis of Literary Style - A Review *J.R.Statist.Soc., 148, Part 4*, 328-341.
[10] Lebart L. (1982). Exploratory Analysis of Large Sparse Matrices, with Application to Textual Data. *COMPSTAT*, Physica Verlag, 67-76.
[11] Lebart L., A.Salem, E. Berry. (1991). Recent Development in the Statistical Processing of Textual Data. in *Applied Stoch. Model and Data Analysis*, 7, 47-62, Wiley.
[12] Lebart L., Morineau A., Warwick. (1984). *Multivariate Descriptive Statistical Analysis*, Wiley, New York.
[13] Salem A.(1987). *Pratique des Segments Répétés, Essai de Statistique Textuelle*, Klincksieck, Paris.