Correspondence Analysis and Classification

Gilbert Saporta

Conservatoire National des Arts et Métiers, Paris

saporta@cnam.fr

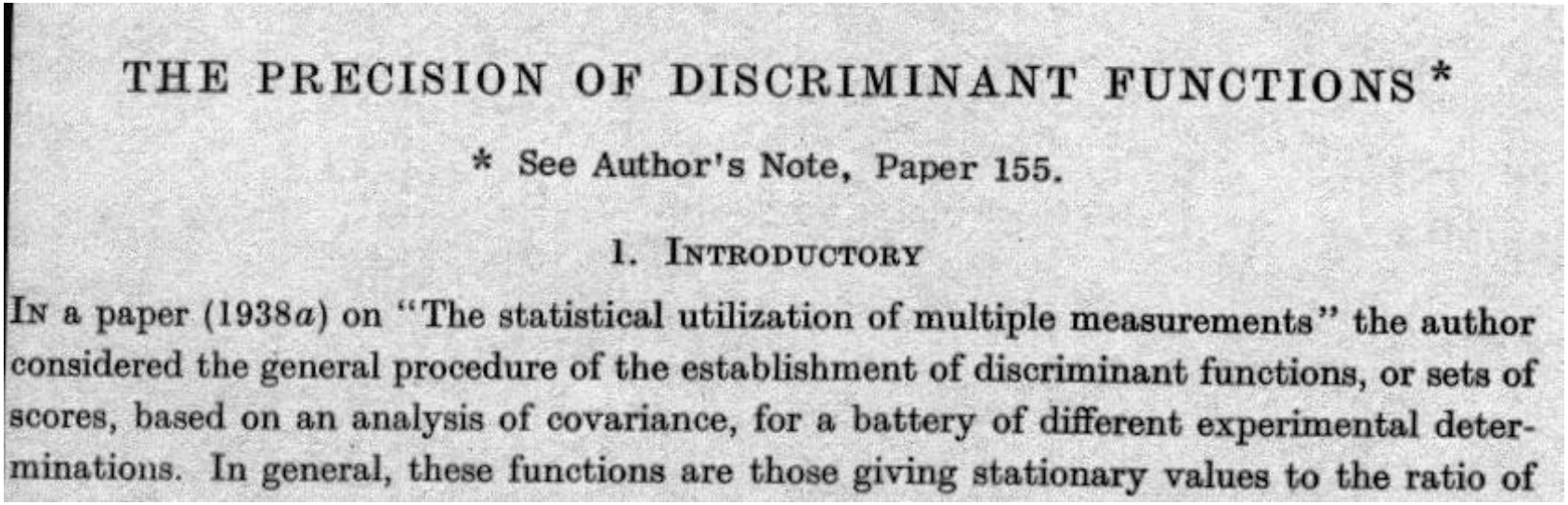
- Classification
  - supervised learning
  - discrimination
  - pattern recognition
- Variables
  - Y response categorical variable (k categories)
  - $X_1,\ldots,X_p$ categorical predictors
- k=2

Motivation

- Credit scoring: risk assessment of loans
- Logistic regression seems to win against discrimination, especially for categorical predictors
- Disqual methodology based on a combination of MCA and Fisher's LDA

A bit of (pre)history

- Fisher 1940
  - Only one predictor
  - Identical to correspondence analysis
  - « Scores » were introduced



THE PRECISION OF DISCRIMINANT FUNCTIONS *

* See Author's Note, Paper 155.

1. INTRODUCTORY

IN a paper (1938*a*) on "The statistical utilization of multiple measurements" the author considered the general procedure of the establishment of discriminant functions, or sets of scores, based on an analysis of covariance, for a battery of different experimental determinations. In general, these functions are those giving stationary values to the ratio of

For example, in a contingency table individuals are cross classified in two categories, such as eye colour and hair colour, as in the following example (Tocher's data for Caithness compiled by K. Maung of the Galton Laboratory).

Eye colour	Hair colour					
	Fair	Red	Medium	Dark	Black	Total
Blue	326	38	241	110	3	718
Light	688	116	584	188	4	1580
Medium	343	84	909	412	26	1774
Dark	98	48	403	681	85	1315
Total	1455	286	2137	1391	118	5387

Variation among the four eye colours may be regarded as due to variations in three variates defined conveniently in some such way as the following:

Eye colour	x_1	x_2	x_3
Blue	0	0	0
Light	1	0	0
Medium	0	1	0
Dark	0	0	1

We may then ask for what eye colour scores, i.e. for what linear function of x_1, x_2, x_3, are the five hair colour classes most distinct. The answer may be found in a variety of ways. For example, by starting with arbitrarily chosen scores for eye colour, determining from these average scores for hair colour, and using these latter to find new scores for eye colour.

Apart from a contraction of scale by a factor R^2 for each completed cycle, this form tends to a limit, and yields scores such as the following:

Eye colour	x	Hair colour	y
Light	−0·9873	Fair	−1·2187
Blue	−0·8968	Red	−0·5226
Medium	0·0753	Medium	−0·0941
Dark	1·5743	Dark	1·3189
		Black	2·4518

The particular values given above have been standardized so as to have mean values zero, and mean square deviations unity. In the sample from which they are derived each score has a linear regression on the other, the regression coefficient being 0·44627; this is, of course, equal to the correlation coefficient between the two scores regarded as variates. Hotelling has called pairs of functions of this kind canonical components. It may be noticed that no assumption is introducéd as to the order of the classes of each category. In Tocher's schedule Light eyes come between Blue and Medium, but the discriminant function puts Blue between Medium and Light, though near the latter.

General case: p predictors

- Optimal scaling approach:
  - Allot partial scores to predictor categories in order to maximize Mahalanobis distance in $\mathbb{R}^p$
  - Ie : transform qualitative variables into numerical ones

- Perform a discriminant analysis where categorical variables are replaced by indicator variables

$$X = \left(\begin{array}{ccc|cc} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{array}\right)$$

Categorisation: a way towards non-linear classification

- Score $S = \sum_{j=1}^{p} \varphi_j(X_j)$
- φ_j step-functions
- Useful for mixed type data
- Easy to add interactions: $X_j * X_k$

- X not of full rank: rank(X)=Σm_i-p
  - Classical solution (GLM or logistic regression): discard one indicator variable for each predictor
  - Disqual (Saporta, 1975): LDA performed on a selection of components of MCA of X. Similar to PCR. Components selected in an expert way according to 2 criteria: inertia and correlation with the response

scorecard

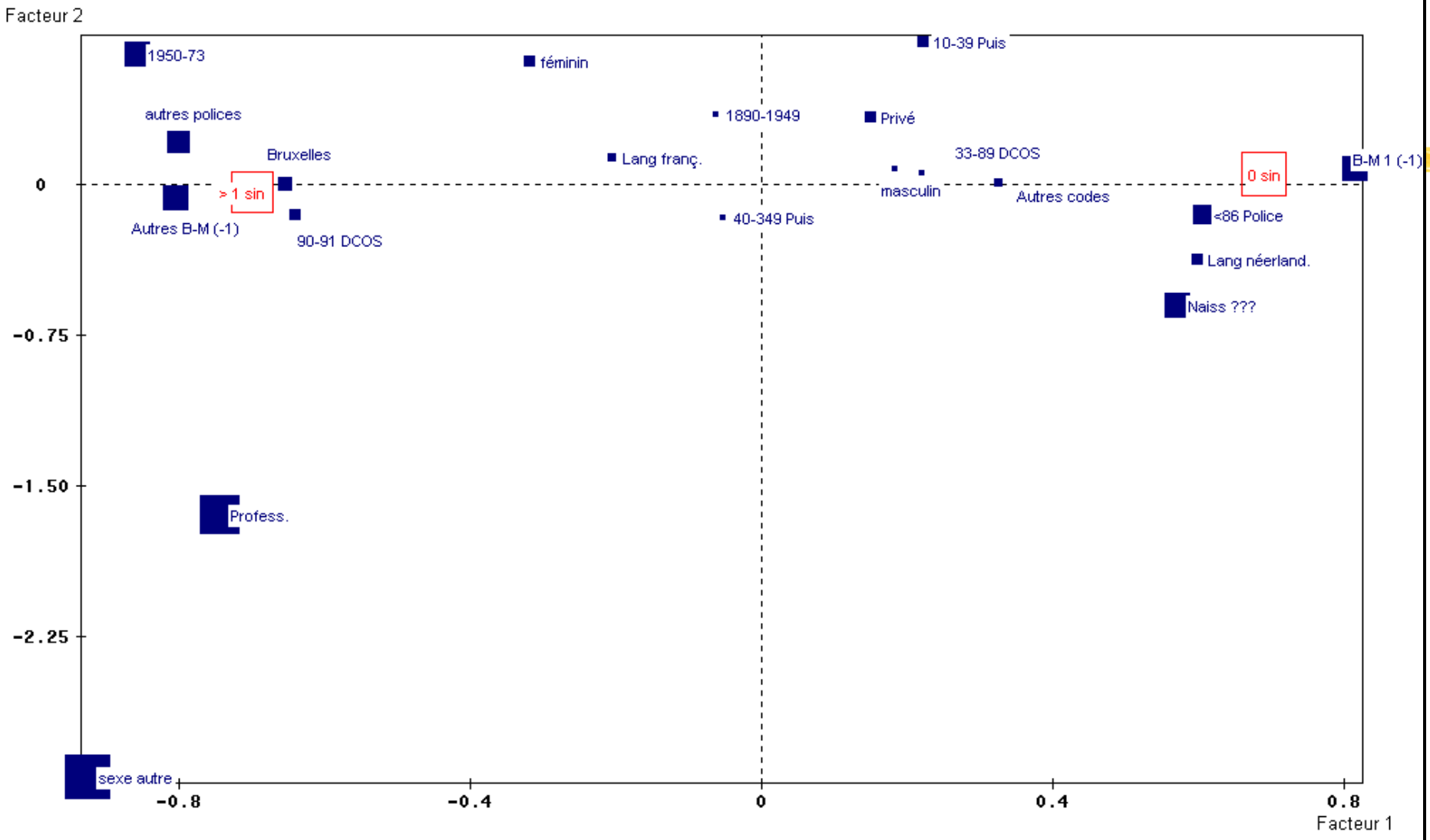
- Transition formulas: a linear combination of row (statistical units) coordinates is given by a linear combination of column (categories) coordinates

Score $s = \sum_{j=1}^{k} d_j \, z^j$ $\qquad s = \sum_{j=1}^{k} d_j X u^j = X \underbrace{\sum_{j=1}^{k} d_j u^j}_{\text{score-card}}$

$$\begin{pmatrix} . \\ d_j \\ . \end{pmatrix} = V^{-1}(g_1 - g_2) = \begin{pmatrix} . \\ \dfrac{\bar{z}_1^j - \bar{z}_2^j}{V(z^j)} \\ . \end{pmatrix}$$

An insurance example (SPAD data set)

- 1106 belgian automobile insurance contracts :
- 2 groups: « 1 good », « 2 bad »
- 9 predictors: 20 categories
  - Use type(2), gender(3), language (2), agegroup (3), region (2), bonus-malus (2), horsepower (2), duration (2), age of vehicle (2)


Facteur 2
1950-73
féminin
10-39 Puis
autres polices
1890-1949
Privé
Bruxelles
Lang franç.
33-89 DCOS
B-M 1 (-1)
0 sin
> 1 sin
masculin
Autres codes
0
Autres B-M (-1)
40-349 Puis
<86 Police
90-91 DCOS
Lang néerland.
Naiss ???
-0.75
-1.50
Profess.
-2.25
sexe autre
-0.8
-0.4
0
0.4
0.8
Facteur 1

Fisher's LDA

FACTEURS	CORRELATIONS	COEFFICIENTS
1 F 1	0.719	6.9064
2 F 2	0.055	0.7149
3 F 3	-0.078	-0.8211
4 F 4	-0.030	-0.4615
5 F 5	0.083	1.2581
6 F 6	0.064	1.0274
7 F 7	-0.001	0.2169
8 F 8	0.090	1.3133
9 F 9	-0.074	-1.1383
10 F 10	-0.150	-3.3193
11 F 11	-0.056	-1.4830
CONSTANTE		0.093575

R2 = 0.57923 F = 91.35686
D2 = 5.49176 T2 = 1018.69159

Score= 6.90 F1 - 0.82 F3 + 1.25 F5 + 1.31 F8 - 1.13 F9 - 3.31 F10 + 0.094.

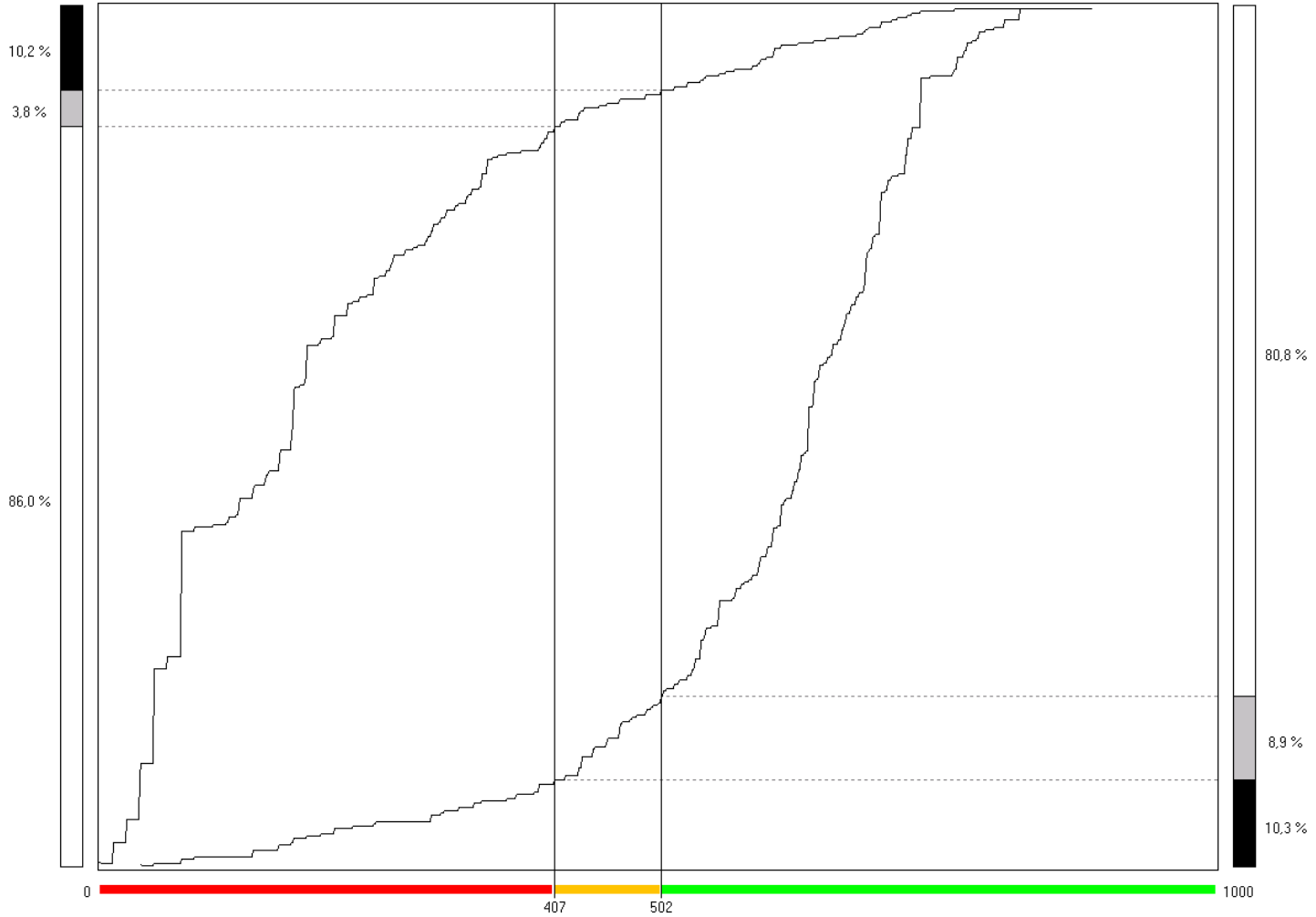
Scorecard

2	CUS1	Profess.	-6.6582
2	CUS2	Privé	1.3374
4	MASC	masculin	0.5201
4	FEMI	féminin	1.9830
4	SOCI	sexe autre	-12.9013
5	FRAN	Lang franç.	-0.7244
5	NEER	Lang néerland.	2.1168
24	AGE1	1890-1949	-4.1563
24	AGE2	1950-73	-15.7429
24	AGE?	Naiss ???	12.3298
25	COD1	Bruxelles	-8.1546
25	COD2	Autres codes	4.0497
26	BM01	B-M 1 (-1)	17.7667
26	BM02	Autres B-M (-1)	-17.5115
27	P86	<86 Police	1.8717
27	P87	autres polices	-2.4682
28	PU01	10-39 Puis	3.6240
28	PU04	40-349 Puis	-0.8846
29	DC01	33-89 DCOS	2.7019
29	DC02	90-91 DCOS	-7.8576
		CONSTANTE	0.095862

COEFFICIENTS DES FONCTIONS DISCRIMINANTE ET SCORE

IDEN	LIBELLES	COEFFICIENTS FONCTION DISCRIMINANTE	COEFFICIENTS TRANSFORMES (SCORE)
2	. Code usage - CUSA 5-6		
CUS1	- Profess.	6.658	66.25
CUS2	- Privé	-1.337	0.00
4	. Sexe - SEXE 11-12		
MASC	- masculin	-0.520	12.12
FEMI	- féminin	-1.983	0.00
SOCI	- sexe autre	12.901	123.33
5	. Code Langue - CLAN 14-15		
FRAN	- Lang franç.	0.724	23.54
NEER	- Lang néerland.	-2.117	0.00
24	. Age de l'assuré (3 mod) - DNAI 8-9		
AGE1	- 1890-1949	4.156	136.61
AGE2	- 1950-73	15.743	232.61
AGE?	- Naiss ???	-12.330	0.00
25	. Code postal souscripteur (2 mod) - POSS2 17-18		
COD1	- Bruxelles	8.155	101.13
COD2	- Autres codes	-4.050	0.00
26	. Bonus-malus Année -1 (2 mod) - GBM1		
BM01	- B-M 1 (-1)	-17.767	0.00
BM02	- Autres B-M (-1)	17.512	292.32
27	. Date effet Police (2 mod) - DPEP 26-27		
P86	- <86 Police	-1.872	0.00
P87	- autres polices	2.468	35.96
28	. Puissance du véhicule (2 mod) - PUIS 32-33		
PU01	- 10-39 Puis	-3.624	0.00
PU04	- 40-349 Puis	0.885	37.36
29	. Année de construction du véhicule (2 mod) - DCOS 38-39		
DC01	- 33-89 DCOS	-2.702	0.00
DC02	- 90-91 DCOS	7.858	87.50

Both CDFs of score function



Why logistic regression?

$$\pi(x) = P(Y = 1 / X = x) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}$$

$$score = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

- In favour among econometricians
  - Looks more « scientific » than data analysis: model instead EDA and geometry, maximum likelihood estimation, standard errors, interpretation of coefficients as odd-ratios
  - Software procedure allows categorical predictors, without manipulating indicator variables

- But:
  - Degeneracies in case of perfect separation
  - Conditional likelihood, asymptotics
  - Standard errors may be computed by bootstrap in LDA
  - In practice: *« It is generally felt that logistic regression is a safer, more robust bet than the LDA model, relying on fewer assumptions . It is our experience that the models give very similar results , even when LDA is used in inappropriately, such as with qualitative variables. » Hastie and al.(2001)*

A need for validation

- A model should be choosen according to its performance, not to ideology!
- Predicting capability or generalisation for new data. Comparisons should be made on test sample. Forecast the future and not the past…

Dimension reduction and generalisation

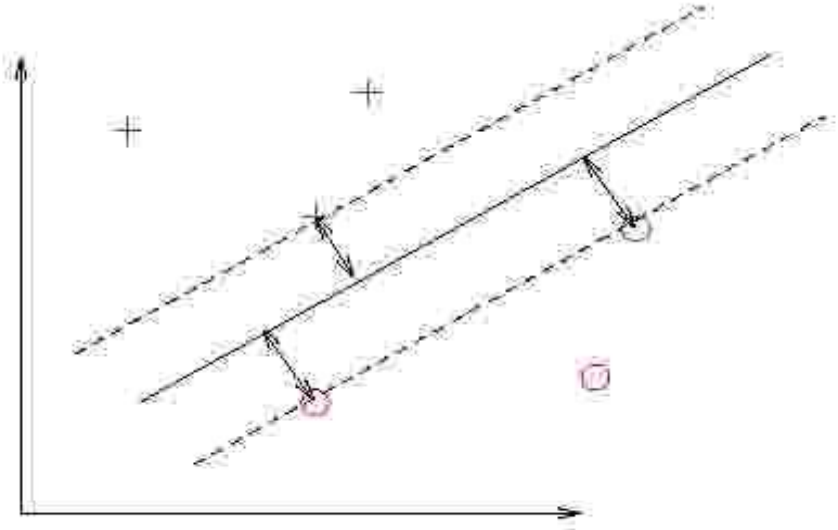
- Components selection should improve performance because of a lower complexity
- VC dimension: a new version of Ockham's razor

Empirical risk and VC dimension

- Vapnik's inequality
  - With probability 1-q

$$R < R_{\text{emp}} + \sqrt{\frac{h\left(\ln\left(2n/h\right)+1\right)-\ln q/4}{n}}$$

  - h : VC dimension different from the dimension of the space



$$h \leq \frac{R^2}{C^2} \quad \text{where } \|x\| \leq R$$

R , radius of the sphere containing all observations decreases when one discards principal axes.

Potential improvements

- Logistic regression using selected MCA components instead of raw variables
- MCA components are computed without taking into account the response:
  - Non-symmetric factor analysis or PLS regression

PLS discriminant analysis

- Y with 2 values (1,2) or (-1,+1)
- Tucker's criterium

$$\max \quad (\mathrm{cov}(y; Xw))^2$$

$$(\mathrm{cov}(y; Xw))^2 = r^2(y; Xw).V(Xw).V(y)$$

- First PLS component:

$$w_j = \frac{\operatorname{cov}(y; x_j)}{\sqrt{(\sum \operatorname{cov}(y; x_j)^2}} \qquad t = \sum w_j x_j$$

- Following component t_2 :

$$y = c_1 t_1 + y_1 \qquad x_j = a_j t_1 + x_{1j}$$
$$t_2 = \sum w_{2j} x_{1j} \qquad \max \operatorname{cov}(y_1; t_2)$$

- Stopping rule: crossvalidation
- Only univariate regressions

PLS and barycentric discrimination

- First PLS component: univariate regression onto all indicator variables
- For any qualitative variable, its indicators are orthogonal
- Getting the first PLS component comes down to p PLS regression performed separately

- Each PLS of Y against indicators of Xj is equivalent to OLS regression (Y should be standardised, not X, and no intercept)
- CA of a contingency table with 2 rows give only one factor
- PLS with one component is equivalent to CA of the concatenation of the contingency tables crossing Y with the Xj

Barycentric discrimination

(A. Leclerc 1976):

		good	bad
1	cusag1	29	96
2	cusag2	344	272
3	sexe1	288	253
4	sexe2	76	78
5	sexe3	9	37
6	clang1	250	295
7	clang2	123	73
8	age3m1	118	99
9	age3m2	40	163
10	age3m3	215	106
11	cpost2m1	75	172
12	cpost2m2	298	196
13	bm2m_11	298	59
14	bm2m_12	75	309
15	puis2m1	91	47
16	puis2m2	282	321
17	dpoli2m1	277	137
18	dpoli2m2	96	231

Only one factor: a pocket computer is enough

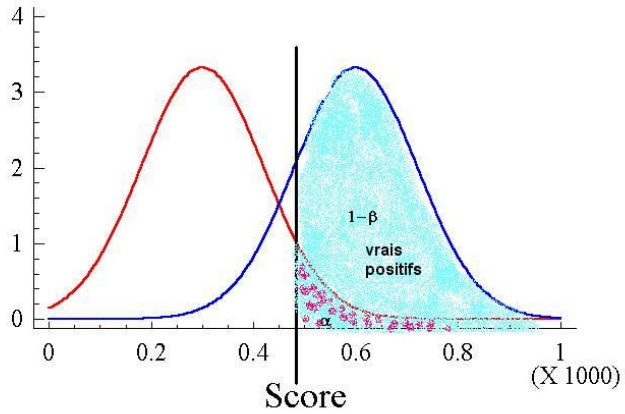
- Ordination of categories along an axis
- If conventionnally group 1 is at the origin, group 2 at 1, a category j of a predictor has a coordinate equal to $n_{j2}/n_{j,}$ conditional frequency of G2/j
- The score of an individual is proportional to the sum of the coordinates of its categories.

- G1 ——————————— G2
- Optimal if independent predictors

Numerical experiments

- Insurance data set split into learning (752) and test sample (356) ten times
- Five methods:
  - discrimination with selection of MCA factors
  - Logistic regression on raw data
  - Logistic regression on selected factors
  - PLS regression with CV choice of the number of components
  - Barycentric discrimination

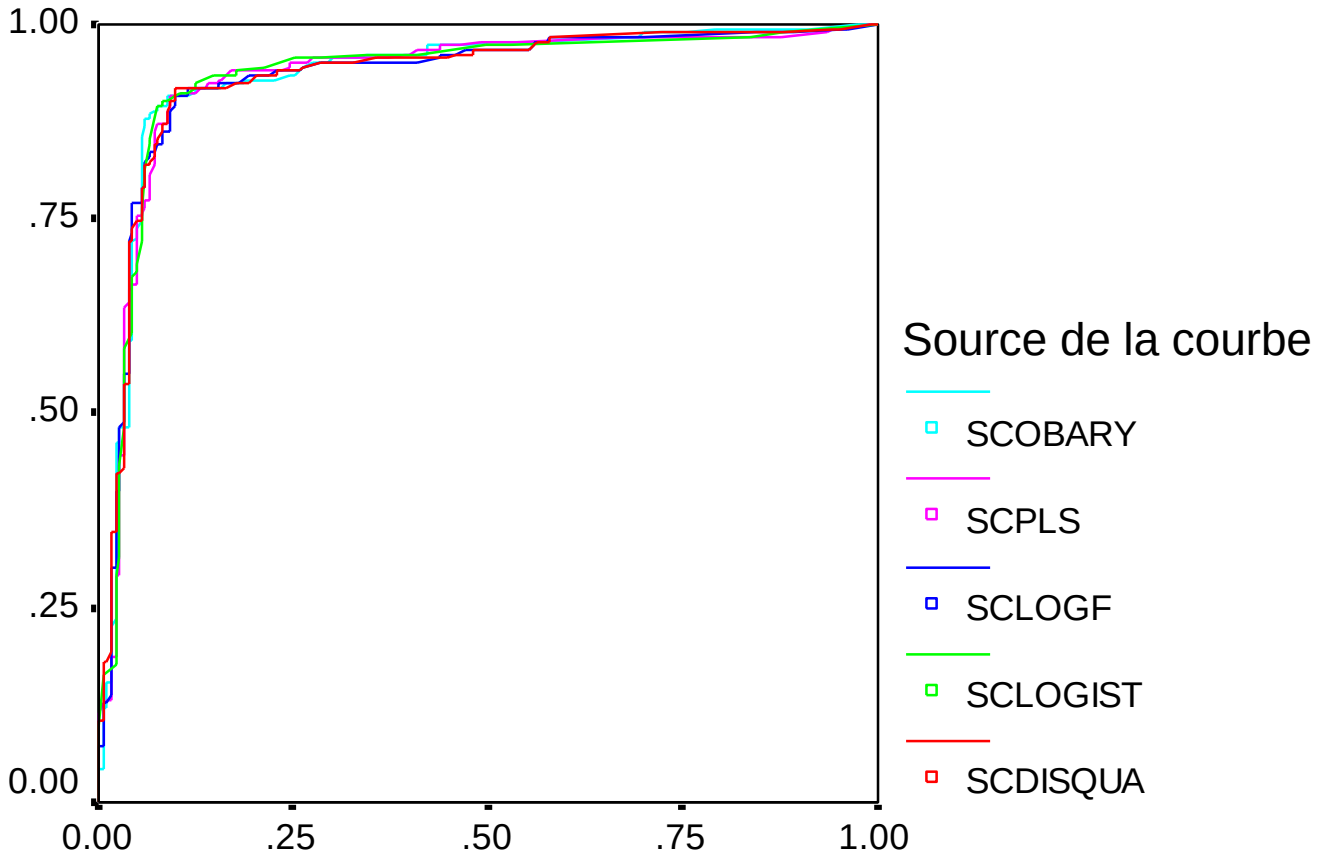
- Five scores computed for the test sample
- Comparison of scores through ROC curve (invariant with any monotonous transformation, takes into account all possible thresholds) : sensitivity (1-β) against α 1-specificity.



Pearson Correlation Coefficients, N = 365

	scdisc	sclogb	sclogf	scpls	fact
scdisc	1.00000	0.96876	0.99885	0.98032	0.97464
sclogb	0.96876	1.00000	0.96821	0.99070	0.96218
sclogf	0.99885	0.96821	1.00000	0.97996	0.97597
scpls	0.98032	0.99070	0.97996	1.00000	0.97735
fact	0.97464	0.96218	0.97597	0.97735	1.00000

Courbe ROC



1 - Spécificité

Les segments diagonaux sont générés par des liaisons.

Area under ROC curve

score	area
SCDISQUA	.934
SCLOGIST	.933
SCLOGF	.932
SCPLS	.933
SCOBARY	.935

Area under the ROC curve:
P(X2>X1)
estimated by the proportion of concordant pairs
closely related to Mann-Whitney statistic

Ridge regression

$$\hat{\beta} = (X'X + kI)^{-1} X'y$$

$$\min \left\| y - X\beta \right\|^2 \quad \text{with} \quad \left\| \beta \right\|^2 < d^2$$

$$X = U \Lambda^{1/2} V' \quad X\hat{\beta}_{ols} = UU'y$$

$$\hat{y} = X\hat{\beta}_{ridge} = X(X'X + kI)^{-1} X'y =$$

$$= U \Lambda^{1/2} (\Lambda + kI)^{-1} \Lambda^{1/2} U'y = \sum_j u_j \frac{\lambda_j}{\lambda_j + k} u_j' y$$

Ridge regression shrinks the eigenvalues, Disqual discards some

- Insurance example:
  - k optimized by crossvalidation with 10 subsamples
  - optimal value of very low: k=0.1
  - no improvement

Concluding remarks

- On a real example:
- Logistic regression does not show any superiority
- MCA with component selection does as well as component oriented methods like PLS. They should be more robust for they avoid overfitting
- Surprisingly barycentric discrimination works well, even if predictors are not independent. See « Lancaster independence model »

- Developments
  - Optimal scaling through direct optimization of the area under the ROC curve (or of the lift chart)
  - Maximum margin hyperplane and SVM
- Explanatory power of MCA which provides an useful description of the data through its ability to capture the structure of the data
- Reducing the dimension of the space always useful!