Correspondence analysis and classification

Gilbert Saporta, Ndeye Niang Keita

CORRESPONDENCE ANALYSIS AND CLASSIFICATION

Gilbert Saporta, Ndeye Niang

The use of correspondence analysis for discrimination purposes goes back to the "prehistory" of data analysis (Fisher, 1940) where one looks for the optimal scaling of categories of a variable X in order to predict a categorical variable Y. When there are several categorical predictors a commonly used technique consists in a two step analysis: multiple correspondence on the predictors set, followed by a discriminant analysis using factor coordinates as numerical predictors (Bouroche and al.,1977).

However in banking applications (credit scoring) logistic regression seems to be more and more used instead of discriminant analysis when predictors are categorical. One of the reasons advocated in favour of logistic regression, is that it gives a probabilistic model and it is often claimed among econometricians that the theoretical basis is more solid, but this is arguable. No doubt also that this tendency is due to the the flexibility of logistic regression software. However it could be easily proved that discarding non informative eigenvectors gives more robust results than direct logistic regression, for it is a regularisation technique similar to Principal Component Regression (Hastie and al. 2001). For two class discrimination, we present a combination of logistic regression and correspondence analysis.

Since factor coordinates are derived without taking into account the response variable, one may use PLS regression which is related to barycentric discrimination (Celeux & Nakache 1994) and to nonsymmetric correspondence analysis (Verde & Palumbo 1996).

1.Introduction

1.1 A bit of (pre)history

In a famous paper, R.A. Fisher (1940) derived the equations of correspondence analysis when solving a particular problem of discrimination on the data from Tocher compiled by Maung: the data is a cross classification of two categorical variables: hair and eye colours of 5387 scottish children. The problem addressed by Fisher was to derive a linear combination of the indicator variables of eye colours giving the best discrimination between the 5 classes of hair colour (table 1).

	Hair colour					
Eye colour	***Fair***	***Red***	***Medium***	***Dark***	***Black***	**Total**
Blue	326	38	241	110	3	719
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

Table 1: Hair and eye colour of scottish chidren

A linear combination of indicator variables leads to score the corresponding categories of eye colour: it comes down to transform a categorical variable into a discrete quantitative variable by allotting scores to each category. It is the beginning of a long sery of works about optimal scaling (see Young 1981).
In other words Fisher performs a canonical discriminant analysis between two sets of variables: the 5 indicators of hair colour on one hand, the 4 indicator variables of the eye colour on the other hand. Actually Fisher used only the indicator variables of the last three eye colours (light, medium, dark) discarding the first indicator of eye colour blue, in order to avoid the trivial solution.
It is well known that the optimal solution is given by the first factor of correspondence analysis of the contingency table: the optimal scores are the coordinates of the categories along the first axis. In his solution, Fisher standardized the scores so as to have zero mean and unit variance when weighted by the marginal frequencies.

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

Table 2: Eye and hair colour scores

The algorithm of successive averages given by Fisher *("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")* is an alternated least squares one and may be viewed as the ancestor of Gifi's (1990) homogeneity analysis and of Nishisato's (1980) dual scaling.
Was Fisher the father of correspondence analysis? Despite the fact that he derived the eigenequation of CA, one cannot say so, for he used only the first eigenvector, preventing the use of graphical displays, which characterizes CA as an exploratory data analysis technique.

2. Linear methods for classification

Let us consider the case of two classes and *p* numerical predictors. The two main techniques are Fisher's linear discriminant analysis (LDA) and logistic regression.

2.1 Fisher's linear discriminant function (LDF)

The linear combination *u* of the *p* variables which maximizes the between to within variance ratio $\frac{\mathbf{u'Bu}}{\mathbf{u'Wu}}$ is given by :

$$\mathbf{u} = \mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) \qquad (1)$$

It corresponds to a linear frontier in the unit space given by the mediator hyperplane separating the two centroids (according to Mahalanobis metric $\mathbf{W}^{-1}$). It is well known that apart from a multiplicative constant, Fisher's LDF is the OLS estimate of $\boldsymbol{\beta}$ in the model $\mathbf{y}=\mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ where **y** takes only two different values, one for each group. Fisher's score of unit **e** can be defined by :

$$S(\mathbf{e}) = (\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{W}^{-1}\mathbf{e} - \tfrac{1}{2}(\mathbf{g}_1'\mathbf{W}^{-1}\mathbf{g}_1 - \mathbf{g}_2'\mathbf{W}^{-1}\mathbf{g}_2) \qquad (2)$$

There is a probabilistic model leading to this result: if the conditional distributions of the p variables in each group are normally distributed with the same covariance matrix $N_p(\mu_i, \Sigma)$ and equal prior probabilities, then the posterior probability for an observation e coming from group 1 is:

$$P(G_1 / e) = \frac{\exp(S(e))}{1+\exp(S(e))} = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}} \qquad (3)$$

Modifying priors changes only the constant term in the score function.

2.2 Logistic regression

In this model, one uses formula (3) as an hypothesis and not as a consequence. The coefficients β_j are estimated by conditional maximum likelihood (ML), while in discriminant analysis they are estimated by least-squares (which are also the unconditional ML estimates in the normal case with equal covariance matrices). In many cases one has often observed that both solutions are very close (see Hastie & al. 2001). Logistic regression is very popular in biostatistics and econometrics since the β_j are related to odds-ratio.
The exponent $\beta_0 + \beta_1 x_1 + ... + \beta_p x_p$ is also used as a score function.

2.3 About the notion of score

In some applications, there is a strict decision rule with a threshold s_0 that says that if $S(e) > s_0$ then e is classified in group 1. But in many other applications one just uses a score as a rating of the risk to be a member of one group, and any monotonic increasing transformation of S is also a score. Let us remark that in this sense the probability $P(G_1 / e)$ is also a score with a range equal to 1.

3. The "*Disqual*" methodology

3.1 Categorical discriminant analysis

Classifying observations described by categorical predictors into one out of k classes has long been done by using models derived from the multinomial or log-linear model, see Goldstein & Dillon (1978) and Celeux & Nakache (1994) for a very comprehensive book. However these models suffer from some curse of dimensionality and are difficult to apply when the number of predictors is large.

An other way of dealing with categorical predictors consists in transforming them into numerical ones by giving numerical values to the categories in an "optimal" way. "Optimal" means that the discrimination which will be done afterwards will maximize some criterium. In the case of two groups, the easiest criterium to optimize is the Mahalanobis distance.

Since transforming qualitative variables into discrete numerical ones comes down to define linear combinations of the indicator variables of the categories, linear discriminant analysis for categorical predictors is a discriminant analysis where predictors are indicator variables.

The scores allotted to the categories define what is called a "scorecard" in credit scoring applications (see Thomas & al. 2002): ie the coefficients of indicator variables.

Let X_1, X_2,..., X_Q be the predictors with J_1,J_2,..., J_Q categories, then one has to do a canonical analysis between the matrix **Y** of the indicator variables of groups and the disjunctive matrix **Z** of

the *Q* predictors:

$$\mathbf{Z} = \begin{pmatrix} 0\,1\,0\,|10|... \\ 0\,0\,1\,|0\,1|\,... \\ ..... \\ 1\,0\,0| \end{pmatrix}$$

However the within-group covariance matrix **W** is not of full rank, since the sum of indicator variables for each predictor is equal to 1, which implies that there is an infinite number of solutions in terms of coefficients (or category scores). Like in the general linear model, one way of getting a solution consists in discarding one indicator variable for each predictor, or in an equivalent way to impose a zero score for this category (usually the last one for most statistical softwares). This is also the solution used in logistic regression for categorical predictors.

3.2 Discrimination with MCA factors

Another solution, named *Disqual*, for finding a scorecard has been proposed by Saporta (1976) and was widely used (at least in France) for credit scoring, see Bouroche & Saporta (1988). It consists in two main steps:

a. a multiple correspondence analysis is performed on the array of predictors **Z** with the class variable as a supplementary one.
b. a linear discriminant analysis (Fisher's LDF for two groups) is done using factor coordinates as predictors.

Since MCA is closely related to PCA, *Disqual* is very close to principal components regression.

The overall score *s* given by Fisher's LDF is a linear combination of the coordinates $s = \sum_{j=1}^{J-Q} d_j \, f^j$ which is not easy to use for new observations. However the transition formulas of MCA allows to write *s* as a sum of the partial scores (the scorecard) of all categories:

$$s = \sum_{j=1}^{J-Q} d_j \mathrm{Z}\mathrm{u}^j = \mathrm{Z} \underbrace{\sum_{j=1}^{J-Q} d_j \mathrm{u}^j}_{\text{score-card}} \qquad (4)$$

where u^j is the *j*th column-factor. The scorecard is a linear combination of the coordinates of the categories along the various axis of MCA, where the coefficients d_j are given by Fisher's formula (1). We use here $\mathbf{V}^{-1}$ instead of $\mathbf{W}^{-1}$, because **V** (the covariance matrix of factor components of MCA) is diagonal: components are uncorrelated.

$$\begin{pmatrix} . \\ d_j \\ . \end{pmatrix} = \mathrm{V}^{-1}\left(\mathrm{g}_1 - \mathrm{g}_2\right) = \begin{pmatrix} . \\ \dfrac{\overline{f}_1^{\,j} - \overline{f}_2^{\,j}}{V(f^j)} \\ . \end{pmatrix} \qquad (5)$$

If the components are correctly standardised, their variances are equal to the eigenvalue or inertia, but it is not always the case: it is software dependent.

3.3 Factor selection

Using all the factors ($J - Q = \sum_{i=1}^{Q} (J_i - 1)$) is equivalent to discard one category for each predictor.

But one of the main interest of the method relies in the possibility of discarding irrelevant factors: factors are computed irrespective of the class variable, some may be not relevant for classification purpose.

Since they are uncorrelated, one has just to use univariate tests of comparison of means: if the class means do not differ significantly on an axis, one can discard it.

What is the interest of using less factors than *J-Q* ? The answer is that doing so gives more robust and reliable prediction for new observations. If the apparent number of parameters (the scorecard coefficients) is still the same, the true degree of freedom decreases and is equal to the number of selected factors.

Statistical learning theory (Vapnik 1998) gives a rationale for that:

Vapnik's inequality (6) states that the true error risk (for new observations from the same distribution) is, with probability 1-*q*, less than the empirical risk (misclassification rate on the learning sample, or resubstitution error rate) plus a quantity depending of the Vapnik-Cervonenkis or VC-dimension *h*:

$$R < R_{\text{emp}} + \sqrt{\frac{h\left(\ln\left(2n/h\right)+1\right) - \ln\left(q/4\right)}{n}} \qquad (6)$$

The VC dimension of a classifier is a measure of its separating power: it is the maximum number of points in a two-class problem which can be perfectly separated in any case.

For instance linear classifiers in $\mathbb{R}^2$ have a VC dimension of 3.

But the VC dimension *h* of a "thick" hyperplane with margin C (C is the distance of the closest point to the hyperplane for completely separated groups) is such that $h \le \frac{\rho^2}{C^2}$ where ρ is the radius of the smallest sphere containing all observations.

If we select only a few factors, we work on a projection of the original data onto a subspace, which lowers ρ. Hence, if the discarded factors were irrelevant for discrimination purposes, R_{emp} does not change, and if the margin remains unchanged, *h* and the bound for the true error risk *R* decrease: we have a better generalization capacity.

4 Alternative methods

Scores may be obtained by various methods, the most widely used being probably logistic regression.

4.1 Logistic regression for categorical predictors

Logistic regression is now extensively used in credit scoring applications, more than discriminant analysis. Our belief is that it is due not only on its specific properties but also to the improvement of specific statistical software for logistic regression. The use of categorical predictors through their transformation into indicator variables is now very easy in most softwares since it is no longer necessary to create indicator variables. One has just to declare predictors as "class variables" and add, if necessary, interactions by simply writing X1*X2. Moreover the stepwise selection is a true variable selection and not a selection of indicator variables. All these features could have been added to discrimination software procedures, but actually it is not the case.

One theoretical drawback of logistic regression is that it uses the full space spanned by the indicator variables, which could be with a high dimensionality and might lead to overfitting.
On the other hand it is well known that logistic regression performs better than linear discriminant analysis when the conditional distribution are not normal or with equal covariances. This is why we propose the following compromise: like in *Disqual* a first step of MCA but followed by a logistic regression with factor selection. The scorecard is then obtained by the same formula (4) presented earlier, the only difference being that the d_j are estimated through a (conditional) maximum likelihood procedure instead of a least squares one.

4.2 PLS regression

Partial least squares regression is a technique alternative to ordinary least squares regression when strong collinearities between predictors is present in the model $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{e}$. PLS looks for a set of uncorrelated linear combination of the predictors $\mathbf{t}_h=\mathbf{X}\mathbf{a}_h$, but unlike principal component regression, PLS components t_h are computed in order to be related to the response **y**.
PLS regression has been proposed as an algorithm by S.Wold but its rationale is better understood in our opinion in terms of maximisation of covariance (Tenenhaus, 1998) ie Tucker's criterium:

$$\max cov^2(\mathbf{y};\mathbf{X}\mathbf{a})$$

Since $cov^2(\mathbf{y};\mathbf{X}\mathbf{w}) = r^2(\mathbf{y};\mathbf{X}\mathbf{w})\, V(\mathbf{X}\mathbf{w})\, V(\mathbf{y})$, maximizing the covariance is a compromise between the explained variance of **X** and the correlation with **y**.
In a classification problem with two classes, one can use PLS regression instead of Fisher's LDF.
However a more symmetric way of dealing with indicator variables, which can be generalized to *k*-groups discrimination, is PLS2 (for multivariate regression) where one maximises $cov^2(\mathbf{Y}\mathbf{b};\mathbf{X}\mathbf{a})$. **Y** is the (n,k) indicator matrix of the groups. If predictors are categorical, **Z** is the disjunctive table. The first PLS component is given by the first eigenvector **a** of $\mathbf{Z'YY'Z}$:

$$\mathbf{Z'YY'Za}=\lambda\mathbf{a}$$

Successive components are obtained by optimizing the Tucker's criterium for residuals after orthogonalization. Usually the number of useful PLS components is chosen by cross-validation.

4.3 Barycentric discrimination

This very simple technique (see Celeux & Nakache 1984) consists in a correspondence analysis of the table $\mathbf{Y'Z}$, which is the concatenation of the Q contingency tables crossing the groups with the Q predictors. If $\mathbf{D_y}$ and $\mathbf{D_z}$ are the diagonal matrices of column frequencies for **Y** and **Z**, the scores for the categories of the predictors are given by the first (and unique if $k=2$) eigenvector of $\frac{1}{Q}\mathbf{D_z^{-1}Z'Y(Y'Y)^{-1}Y'Z}$ since the row-margin diagonal matrix of $\mathbf{Y'Z}$ is $Q\mathbf{D_y}$ and its column-margin is $\mathbf{D_z}$

$$\frac{1}{Q}\mathbf{D_z^{-1}Z'P_yZa}=\lambda\mathbf{a}$$

For a classification into two groups, $\mathbf{Y'Z}$ is a matrix with 2 rows and computations may be done by hand since there is only one axis. The score of a category j is given by the barycenter of $\mathbf{g_1}$ and $\mathbf{g_2}$ weighted by n_{1j} and n_{2j}.
Usually $\mathbf{g_1}$ and $\mathbf{g_2}$ are put at the extremities of the segment [0;1] and the score of an unit is equal to the sum of the conditional probabilities (of being a member of group 2) of its Q categories.
Barycentric discrimination is equivalent to *Disqual* only if the predictors are pairwise independent. It is similar to "naive Bayes classifier" but with an additive score, instead of a multiplicative one.

4.4 Non symmetric correspondence analysis

Proposed by Lauro & d'Ambra (1984) for one variable *X* and used by Verde & Palumbo (1996) for discrimination purpose with *p* predictors X_j, this technique is equivalent to redundancy analysis (Van den Wollenberg 1979) or PCA with instrumental variables (Rao 1964). When the **X** matrix is non-singular (which is not the case for the disjunctive table), the linear combinations of the columns of **X** are the eigenvectors of $\mathbf{V_{11}^{-1}V_{12}V_{21}}$. For categorical predictors we have:

$$\mathbf{D_z^{-1}Z'YY'Za} = \lambda\mathbf{a}$$

When both groups have equal frequencies this comes down to barycentric discrimination and also to the first component of PLS regression.
Following Bougeard & al. (2004) one may derive a continuous set of solutions from *Disqual* or MCA (α=0) to redundancy analysis (α=1) by maximizing:

$$\alpha\, Cor^2(\mathbf{Yb, Xa}) + (1-\alpha)\sum_{j=1}^{p} Cor^2(\mathbf{Xa, X_j a_j})$$

5 A case study

5.1 Data description

The sample consists in 1106 automobile insurees from Belgium observed in 1992 belonging to 2 groups.

- Those without claim n_1=556 (the "good" ones).
- Those with more than one claim (the "bad" ones) n_2=550.

We use here 9 categorical predictors with a total of 20 categories :

- Use type(2),
- Gender (3)
- language (2)
- birth date (3)
- region (2)
- level of bonus-malus (2)
- horsepower (2)
- duration of contract (2)
- year of vehicle construction (2)

There are 3 categories for gender (male, female, others) for an insuree can be a company.

5.2 MCA

The class variable (good or bad) is a supplementary one. MCA gives 11 = 20-9 factors.

NUMBER	EIGEN VALUE	PROPORTION	CUMULATIVE	
1	0.2438	19.95	19.95	**********
2	0.1893	15.49	35.44	*******
3	0.1457	11.92	47.36	*****
4	0.1201	9.82	57.18	****
5	0.1091	8.92	66.11	****
6	0.0999	8.17	74.28	****
7	0.0855	7.00	81.28	***
8	0.0732	5.99	87.26	***
9	0.0573	4.68	91.95	**
10	0.0511	4.18	96.13	**
11	0.0473	3.87	100.00	**

Table 3 : eigenvalue diagram of MCA

The results show that the first factor is very discriminant with a test-value of 23 for the « good bad » variable. When all the predictors are highly related to the group variable, this is often the case. If one could derive a simple rule based only on the first factor, the use of other factors will improve the decision rule.

CATEGORIES		TEST-VALUES					COORDINATES					
IDEN - LABEL	FREQ	1	2	3	4	5	1	2	3	4	5	DISTO.
2 . use type												
USE1 - Profess.	185	11.1	24.5	1.5	-2.1	-3.7	0.74	1.64	0.10	-0.14	-0.25	4.98
USE2 - private	921	-11.1	-24.5	-1.5	2.1	3.7	-0.15	-0.33	-0.02	0.03	0.05	0.20
4 . gender												
MALE - male	787	-9.5	-3.7	-17.4	9.6	16.6	-0.18	-0.07	-0.33	0.18	0.32	0.41
FEMA - female	249	5.7	-10.8	15.6	-8.5	-16.2	0.32	-0.61	0.87	-0.47	-0.90	3.44
COMP - companies	70	8.0	25.5	5.7	-3.4	-3.1	0.93	2.95	0.66	-0.39	-0.36	14.80
5 . Language												
FREN - french	824	11.6	-7.4	14.5	17.9	-1.1	0.20	-0.13	0.26	0.31	-0.02	0.34
FLEM - flemish	282	-11.6	7.4	-14.5	-17.9	1.1	-0.60	0.38	-0.75	-0.92	0.06	2.92
24 . Birth date												
BD1 - 1890-1949 BD	301	1.3	-6.9	-13.9	16.7	-15.1	0.06	-0.34	-0.69	0.82	-0.74	2.67
BD2 - 1950-1973 BD	309	17.7	-13.2	0.2	-13.8	13.2	0.86	-0.64	0.01	-0.67	0.64	2.58
BD? - ???BD	496	-17.1	18.2	12.3	-2.5	1.6	-0.57	0.61	0.41	-0.08	0.05	1.23
25 . Region												
REG1 - Brussels	367	15.3	0.1	15.8	14.7	7.1	0.65	0.00	0.67	0.63	0.30	2.01
REG2 - Other regions	739	-15.3	-0.1	-15.8	-14.7	-7.1	-0.32	0.00	-0.33	-0.31	-0.15	0.50
26 . Level of bonus-malus												
BM01 - B-M +	549	-26.9	-2.3	1.1	3.4	-5.9	-0.82	-0.07	0.03	0.10	-0.18	1.01
BM02 - Other B-M (-1)	557	26.9	2.3	-1.1	-3.4	5.9	0.80	0.07	-0.03	-0.10	0.18	0.99
27 . Duration of contract												
C<86 - <86 contracts	629	-23.2	5.9	10.8	7.0	0.9	-0.61	0.15	0.28	0.18	0.02	0.76
C>87 - other contracts	477	23.2	-5.9	-10.8	-7.0	-0.9	0.80	-0.20	-0.37	-0.24	-0.03	1.32
28 . Horsepower												
HP1 - 10-39 HP	217	-3.6	-11.5	15.5	-12.3	-9.1	-0.22	-0.70	0.94	-0.75	-0.55	4.10
HP2 - 40-349 HP	889	3.6	11.5	-15.5	12.3	9.1	0.05	0.17	-0.23	0.18	0.13	0.24
29 . year of vehicle construction												
YVC1 - 1933-1989 YVC	823	-12.5	-3.0	9.9	-2.0	17.9	-0.22	-0.05	0.17	-0.04	0.32	0.34
YVC2 - 1990-1991 YVC	283	12.5	3.0	-9.9	2.0	-17.9	0.64	0.15	-0.51	0.10	-0.92	2.91
1 . claim												
CLA0 - 0 claim	556	-23.1	-1.5	2.3	1.5	-2.8	-0.69	-0.05	0.07	0.04	-0.08	0.99
CLA1 - > 1 claim	550	23.1	1.5	-2.3	-1.5	2.8	0.70	0.05	-0.07	-0.04	0.08	1.01

Table 4: factor coordinates and test-values for all categories

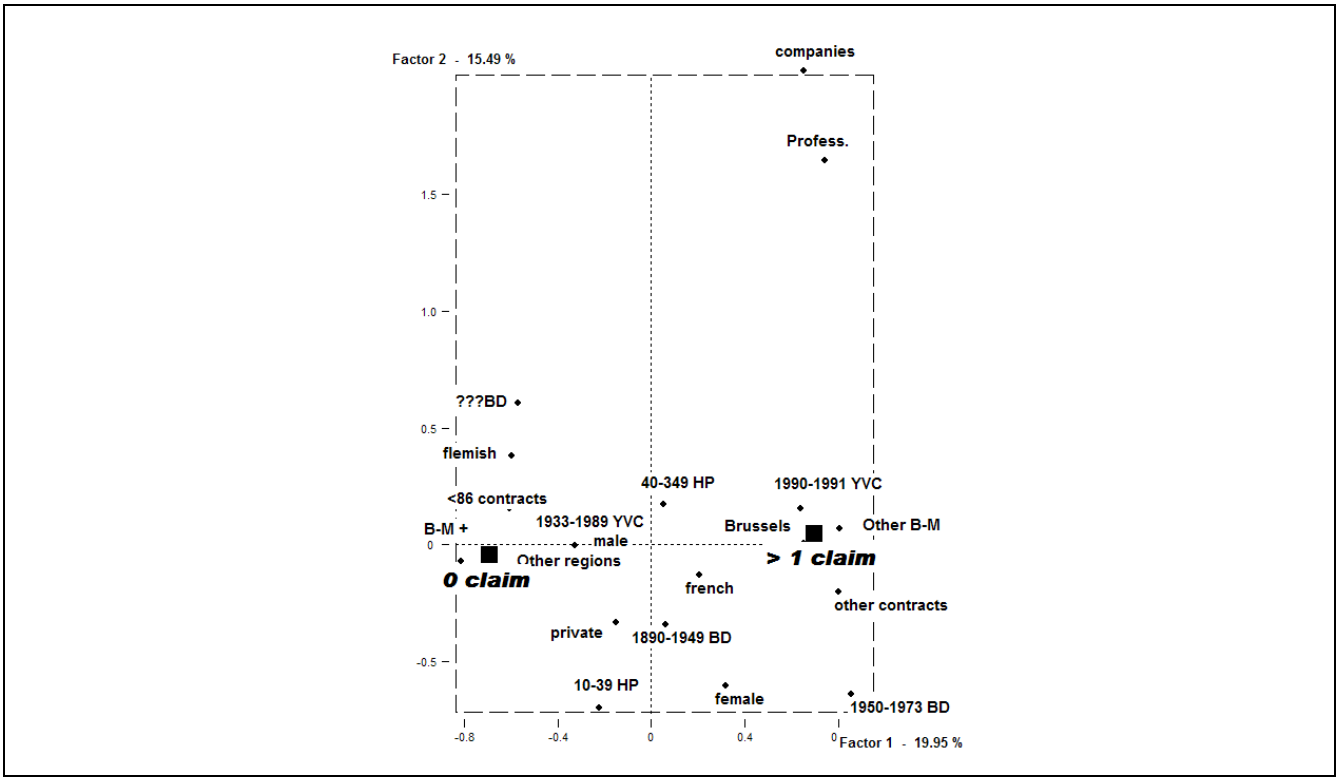

Figure 1: principal plane of MCA

5.3 *Disqual*

The next step consists in performing Fisher's LDA using all factor coordinates as predictors, but all factors are not equally predictive. Since Fisher's LDA is defined up to its sign, we have chosen an orientation such that high values means a "good" driver.

LINEAR DISCRIMINANT FUNCTION

FACTORS	CORRELATIONS WITH L.D.F.	COEFFICIENTS DISC.FUNCTION
F1	**-0.695**	**-6.0525**
F2	*0.046*	*0.4548*
F3	**0.068**	**0.7639**
F4	*0.045*	*0.5530*
F5	**-0.084**	**-1.0876**
F6	**-0.084**	**-1.1369**
F7	*-0.009*	*-0.1270*
F8	**-0.063**	**-1.0064**
F9	**0.079**	**1.4208**
F10	**0.129**	**2.4594**
F11	**0.062**	**1.2324**

Table 5: Fisher's LDF as a combination of factor coordinates

We see that the ranking according to eigenvalues is not the same as the ranking according to the prediction of the class variable. We decide (on the basis of a correlation greater than 0.05) to discard factors F2, F4, F7. Since factors are uncorrelated, the coefficients do not change when some factors are discarded, like in a regression with orthogonal predictors. Thus the score of a statistical unit is given by –0.695F1 +0.068F3 +…+ 0.062F11

Table 6 gives the scorecard with two options: the raw coefficients coming from the direct application of LDA to indicator variables, and the transformed coefficients standardized by a linear transformation such that the score range is [0 ;1000] which is the most used in practice.
The transformed coefficients are obtained in the following way: since the minimal score is equal to –4.577 –2.236 –0.955 -…- 10.222, we first add 4.577 to the scores of both categories of "use type", 2.236 to the three categories of "gender" etc. in order that the miminum score value be zero. The "worst" category of each variable has now a score equal to zero. Each category score is then multiplied by a constant in order that the maximal value of the sum of partial scores be equal to 1000.

CATEGORIES	COEFFICIENTS DISCRIMINANT FUNCTION	TRANSFORMED COEFFICIENTS (SCORE)
2 . Use type		
USE1 - Profess.	-4.577	0.00
USE2 - private	0.919	53.93
4 . Gender		
MALE - male	0.220	24.10
FEMA - female	-0.065	21.30
OTHE - companies	-2.236	0.00
5 . Language		
FREN - French	-0.955	0.00
FLEM - flemish	2.789	36.73
24 . Birth date		
BD1 - 1890-1949 BD	0.285	116.78
BD2 - 1950-1973 BD	-11.616	0.00
BD? - ???BD	7.064	183.30
25 . Region		
REG1 - Brussels	-6.785	0.00
REG2 - Other regions	3.369	99.64
26 . Level of bonus-malus		
BM01 - B-M 1 (-1)	17.522	341.41
BM02 - Others B-M (-1)	-17.271	0.00
27 . Duration of contract		
C<86 - <86 contracts	2.209	50.27
C>87 - others contracts	-2.913	0.00
28 . Horsepower		
HP1 - 10-39 HP	6.211	75.83
HP2 - >40 HP	-1.516	0.00
29 . year of vehicle construction		
YVC1 - 1933-1989 YVC	3.515	134.80
YVC2 - 1990-1991 YVC	-10.222	0.00

Table 6: Score card

The final score is obtained by adding the values corresponding to the categories: eg an insuree with a private use of his vehicle, male, french-speaking etc will get a score of 53.93+24.1+0+ …

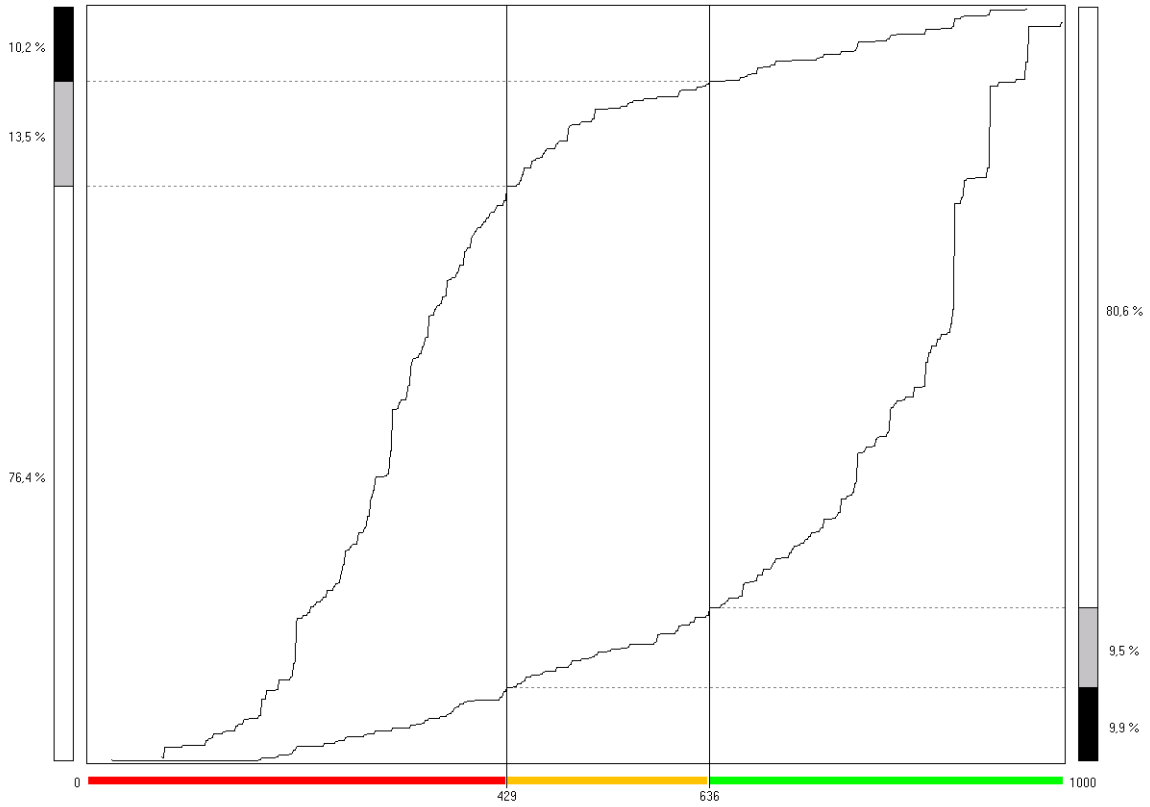

Figure 2: CDF of the score for the two groups

Figure 2 gives the cumulative distribution function of the score for both groups. It is commonly used to derive classification rules, taking into account the two kinds of error risks. Here both risks have been taken equal to 10%: an insuree with a score lower than 429 will be predicted as a “bad” one: more than 75 % of the “bad”ones are detected and 10% of the “good” ones are wrongly considered as “bad”. Conversely 80 % of the “good” insurees have a score higher than 636, and only 10% of the “bad”. The interval [429; 636] is an uncertainty domain.

5.4 A first comparison with logistic regression

We have applied logistic regression to the same data. Table7 gives the coefficients of the logistic score in the column « estimate ». We see that the constraint used here is that the last category of each predictor has a zero score.

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.2498	0.4416	0.3199	0.5716
USE TYPE	Private	1	0.7060	0.2688	6.9000	0.0086
USE TYPE	Profess.	0	0	.	.	.
GENDER	female	1	0.4868	0.4437	1.2039	0.2725
GENDER	male	1	0.4797	0.4074	1.3860	0.2391
GENDER	companies	0	0	.	.	.
LANGUAGE	french.	1	-0.1236	0.2212	0.3124	0.5762
LANGUAGE	flemish	0	0	.	.	.
BIRTH DATE	1890-1949 BD	1	-0.3596	0.2310	2.4218	0.1197
BIRTH DATE	1950-1973 BD	1	-1.6155	0.2512	41.3684	<.0001
BIRTH DATE	??? BD	0	0	.	.	.
REGION	brussels	1	-0.8585	0.2013	18.1904	<.0001
REGION	others regions	0	0	.	.	.
LEVEL OF BM	others B-M(-1)	1	-2.4313	0.1927	159.2260	<.0001
LEVEL OF BM	B-M 1 (-1)	0	0	.	.	.
HORSEPOWER	10-39 HP	1	0.7305	0.2535	8.3037	0.0040
HORSEPOWER	40-349 HP	0	0	.	.	.
DURATION	<86 contracts	1	0.4932	0.2021	5.9536	0.0147
DURATION	others contracts	0	0	.	.	.
YEAR OF VEHICLE	1933-1989 YVC	1	1.3362	0.2095	40.6677	<.0001
YEAR OF VEHICLE	1990-1991 YVC	0	0	.	.	.

Table 7: Scorecard by logistic regression

In this form, it is difficult to compare the results of both methods. If we compute the correlation coefficient between both scores, we find r= 0.97446, which is fairly high, but using a correlation coefficient is right only if the relationship between both scores is linear.

Another way of comparing scores is to compare their ROC curves and AUC (Area under the Roc curve):
ROC curve (see Bamber 1975) synthetizes the performance of a score for any threshold *s*.
Using *s* as a parameter, ROC curve links the probability of being an actual member of G_1 if S>s (true positive) to the probability of being wrongly classified to G_1 (false positive). One of the main properties of ROC curve is that it is invariant by any increasing (not only linear) transformation of *S* . Since the ideal curve is the one which sticks to the edges of the unit square, the favourite measure is given by the area under ROC curve (AUC) which allows to compare several curves (if there is no crossing). Theoretical AUC is equal to the probability of "concordance" : $AUC=P(X_1>X_2)$ when one draws at random two observations independently from both groups. For two samples of n_1 and n_2 observations AUC comes down to Mann-Whitney's U statistic.

Figure 3 shows very close results: logistic regression gives a slightly greater AUC than *Disqual* : 0.908 instead of 0.904, but with a standard error of 0.01 the difference is not significant.

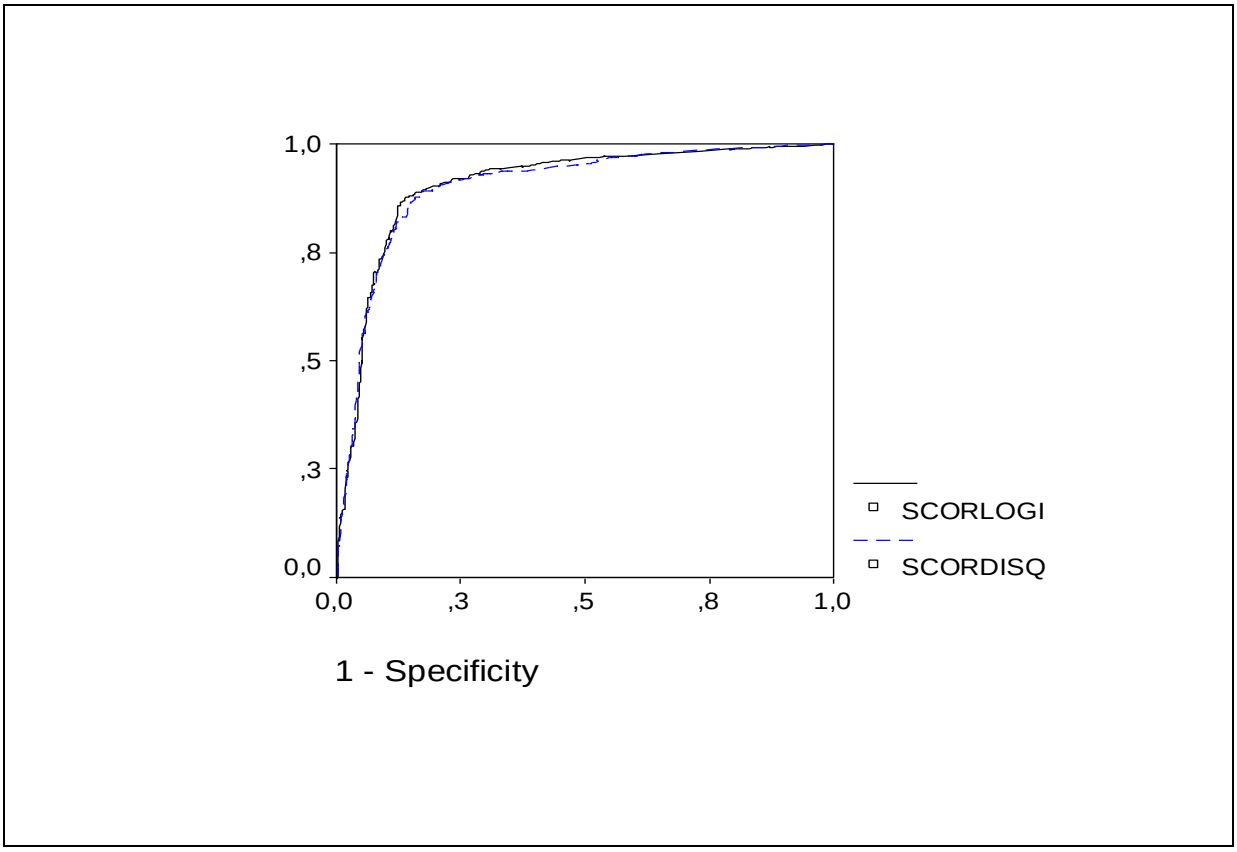

Figure 3: ROC curves

5.5 A more thorough comparison

In the preceding section, the comparison was done with the total sample and may suffer from a resubstitution bias. If we want to compare predicting capabilities of several methods, it is necessary to do so with an independent sample: one has to divide randomly the total sample into two parts: the training set and the test set. In order to avoid a too specific pattern, we did this random split 50 times using a stratified sampling (the strata are the 2 groups) without replacement of 70% for the training sample and 30 % for the test sample.
We used the following five methods:

- *Disqual* with an automatic selection of relevant factors with a probability level 5%
- Logistic regression on raw data at probability
- Logistic regression on MCA factors with automatic selection (probability level 5%)
- PLS regression with cross validation factor selection
- Barycentric discrimination

We notice that the two methods using factor selection do not keep the same set of factors and that there is some variation during the 50 iterations.
Table 8 give the results of factor selection for the first 15 iterations. Factors F1 F5 F6 F9 F10 are selected 15 times by *Disqual*, factor F7 is never selected. Logistic regression is more selective than *Disqual*.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Disqual	F1 F2 F3 F4 F5 F6 F8 F9 F10 F11	F1 F3 F4 F5 F6 F8 F9 F10 F11	F1 F3 F5 F6 F8 F9 F10 F11	F1 F3 F4 F5 F6 F8 F9 F10 F11	F1 F2 F5 F6 F8 F9 F10 F11	F1 F2 F3 F5 F6 F8 F9 F10 F11	F1 F3 F4 F5 F6 F8 F9 F10 F11	F1 F2 F3 F5 F6 F8 F9 F10 F11	F1 F3 F4 F5 F6 F8 F9 F10 F11	F1 F3 F5 F6 F8 F9 F10 F11	F1 F2 F3 F4 F5 F6 F8 F9 F10 	F1 F2 F3 F5 F6 F8 F9 F10 	F1 F3 F5 F6 F8 F9 F10 F11	F1 F3 F5 F6 F8 F9 F10 	F1 F3 F5 F6 F8 F9 F10 F11
NB FAC	10	9	8	9	8	8	8	9	9	7	9	8	8	7	8
Logifact	F1 F3 F6 F8 F9 F10	F1 F3 F5 F6 F8 F9 F10	F1 F3 F6 F9 F10	F1 F3 F5 F6 F8 F9 F10 F11	F1 F6 F9 F10	F1 F3 F6 F9 F10	F1 F3 F5 F6 F8 F10	F1 F3 F5 F6 F8 F9 F10	F1 F3 F6 F8 F9 F10	F1 F3 F5 F6 F9 F10	F1 F2 F5 F6 F8 F10	F1 F3 F5 F6 F8 F9 F10	F1 F3 F5 F6 F9 F10	F1 F3 F5 F6 F8 F10	F1 F3 F5 F6 F9 F10
NB FAC	6	7	5	8	4	5	6	7	6	6	6	7	6	6	6

Table 8: factor selection for discriminant analysis and logistic regression

The performance of each 5 methods was measured by the AUC computed 50 times on the test samples. One remarks in table 8 that the methods based on a selection of MCA factors are more precise (ie with a lower standard deviation) even if the average is slightly not as good.

	Disqual	**Logistic**	**Logist-factor**	**Pls**	**Bary-discrim**
mean	.9024	.9044	.9023	.9035	.9039
std	.0152	.0156	.0146	.0157	.0155
Min	.863	.857	.861	.856	.860
Max	.928	.932	.928	.930	.933

Table 9: AUC on 50 test samples

PLS regression was performed with a crossvalidation choice for the numbers of factors: 4 factors were selected in 42 cases, 3 factors for the other 8 simulations.

Actually the five methods give very similar performance and it is not possible to state the superiority of any one. Barycentric discrimination has very good performance despite its simplicity: an explanation of this surprising fact might be that the nine variables were already the result of an expert selection and have low intercorrelations.

Conclusion

We have advocated multiple correspondence analysis as an intermediate step to derive numerical predictors before applying a linear discriminant analysis. The ability of MCA to recover the data structure explains its efficiency, despite the fact that factors are computed without taking into account the response (group) variable: less factors are necessary if one uses a non-symmetrical analysis like PLS.
The comparison with logistic regression has not shown any systematic superiority of this technique which can be combined with a selection of MCA factors: it is a kind of regularization which lowers the VC dimension.
Moreover the factor space may be used also as a basis for non-linear analysis, which is not possible with barycentric discrimination, and for optimizing other criteria than the Mahalanobis distance, like AUC. The use of factor coordinates gives also a key for applying to categorical data methodologies designed for numerical predictors, such as support vector machines or neural networks.

Software notes

- Multiple correspondence analysis and score functions have been performed with SPAD v5.6 from Decisia (http://www.decisia.fr) (the insurance data set being provided with this software).
- Logistic regression was performed with Proc Logistic from SAS v8.2
- ROC curves and AUC were computed with SPSS v11.5

References

Bamber, D. (1975) The area above the ordinal dominance graph and the area below the receiver operating graph, *Journal of Mathematical Psychology*, **12**, 387-415.

Bougeard S., Nocairi S., Hanafi M., Qannari E.M.,(2004) Description and prediction of categorical variables, in J.Blasius & M.Greenacre eds. ,*Multiple Correspondence Analysis and Related Methods*, Academic Press.

Bouroche, J.M., Saporta G., Tenenhaus M. (1977) Some methods of qualitative data analysis, in *Recent Developments in Statistics*, (J.R.Barra ed.), pp.749-755, Amsterdam: North-Holland

Bouroche, J.M., Saporta G., (1988) Les méthodes et les applications du credit-scoring , *Atti 34° Riunione Scientifica della Sociéta Italiana di Statistica*, p.19-26,

Celeux, G., Nakache, J.P.(1994). *Discrimination sur variables qualitatives*, Paris: Polytechnica

Fisher, R.A. (1940) The precision of discriminant functions, *Annals of Eugenics,* **10**, 422-429.

Gifi, A. (1990) *Non linear multivariate analysis*, New-York: Wiley

Goldstein M., Dillon W.R. (1978) *Discrete discriminant analysis*, New York: Wiley

Hastie T., Tibshirani F., Friedman J. (2001). *The Elements of Statistical Learning.* New-York: Springer

Nishisato S. (1980) *Analysis of categorical data : dual scaling and its applications*, University of Toronto Press

Rao C.R., (1964) The use and interpretation of principal component analysis, *Sankhya*, 26, 329-357

Saporta G. (1976) Discriminant analysis when all the variables are nominal, a stepwise method, *Spring Meeting of the Psychometric Society, Murray-Hill*.

Tenenhaus M. (1998) *La régression PLS*, Paris: Editions Technip

Thomas L.C., Edelman D.B., Crook J.N. (2002) *Credit Scoring and its Applications*, SIAM monographs on Mathematical Modelling and Computation

Van den Wollenberg, A. (1977) Redundancy analysis: an alternative to caninical correlation analysis, *Psychometrika*, 42, 207-219

Vapnik V. (1998) *Statistical Learning Theory*, New York: Wiley

Verde R., Palumbo F. (1996) Analisi fattoriale discriminante non-simmetrica su predittori qualitativi. *Atti del Convegno della XXXVIII Riunione Scientifica della Società Italiana di Statistica*, Rimini.

Young, F.W. (1981) Quantitative analysis of qualitative date. *Psychometrika*, 46, 357-388.

About the authors

Gilbert Saporta is professor and head of the chair of Applied Statistics at CNAM-Paris (Conservatoire National des Arts et Métiers). He is responsible of the data analysis group of CEDRIC, the computer science research team of CNAM. His interests are in applied multivariate analysis, scoring techniques and time dependent data. He has been president of SFdS, the french statistical society.
Chaire de Statistique Appliquée, CNAM, 292 rue Saint Martin, 75141 Paris Cedex 03, France. E-mail : saporta@cnam.fr

Ndeye Niang is assistant professor of statistics at Institut d'Informatique d'Entreprise, the engineering school in computer science of CNAM , and a member of the data analysis group of CEDRIC, the computer science research team of CNAM. Her interests are in multivariate analysis and applications to quality control.
Chaire de Statistique Appliquée, CNAM, 292 rue Saint Martin, 75141 Paris Cedex 03, France. E-mail : niang@cnam.fr

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

Table 1: Hair and eye colour of scottish chidren

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

Table 2: Eye and hair colour scores