Journal of Statistics Applications & Probability
--- An International Journal

An Integrated Approach to Regression Analysis in Multiple Correspondence Analysis and Copula Based Models

Khine Khine Su-Myat[1], Jules J. S. de Tibeiro[1], Pranesh Kumar[2]

[1]*Secteur des Sciences, Université de Moncton, Campus de Shippagan, Shippagan, N.-B., Canada*
[2]*Department of Math. and Statistics, University of Northern British Columbia, Prince George, B.C. Canada*
Email Address: kk.sumyat@gmail.com, kumarp@unbc.ca, jules.de.tibeiro@umoncton.ca

Abstract: In this paper, taking into account the possible development of serious disorders of the proliferation of the plasmatic cells, we focus on a dataset concerning the prediction among a chronic disease which has the higher risk of malignant transformation. The purpose of this paper is to argue in favour of the use of multiple correspondence analysis (MCA) as a powerful exploratory tool for such data. Following usual regression terminology, we refer to the primary variable as the response variable and the others as explanatory or predictive variables. As an alternative, a copula based methodology for prediction modeling and an algorithm to stimulate data are proposed.

Keywords: multiple correspondence analysis, Burt matrix, regression table, regression analysis, barycentric coding, binary logistic regression, copulas.

1 Introduction

Many practical studies adhere to the following scheme: a set of observations I is described by a set of variables Q which can be subdivided into a set of predictive variables J_p and a set of response variables J_r. The problem is to find and explain relationships (causal or not) between the variables of J_p and those of J_r. In general, if J_r is reduced to only one variable, j_r, several traditional methods of prediction are applicable, according to the type of variable j_r and to the types of variables of J_p. For more details, see Rousseau *et al*. [31].

From the clinical point of view, the state of any healthy or sick subject could be completely described by the results of a set of examinations judiciously selected once and for all; the interpretation of the set of results would constitute the diagnosis; the prevision of the later states would be the pronostic. In addition to the traditional checkups, there exist complex sets of examinations which are systematically applied to explore a medical function.

Of primary interest was the possible development of serious plasma cell proliferative disorders, however, the advanced age of many patients makes death from other causes a significant competing risk. Data thus produced may be regarded as a contingency table, where a large amount of data is usually collected on each patient entered, and each column standing for continuous explanatory or response predictors. It is from this point of view our study begins, which relates to the monoclonal gammopathy of undetermined significance (MGUS). These gammopathies correspond to an asymptomatic affection associated with a peak of serum monoclonal immunoglobulin, highlighted at 1% of the 50 year old population, 3% of people over the age of 70 and 10% of the population of more than 80 years.

We refer to a dataset obtained from the private clinic Mayo (USA) where all the 241patients diagnosed, apparently with a benign monoclonal gammapathy before January 1, 1971, were then followed at the beginning of 1992. See (http://mayoresearch.mayo.edu/ mayo/ research/ biostat/ therneau-book.cfm). See also Kyle R.A. [20] and International Myeloma Working Group [19]. For more details about this paper, see Su-Myat [33].

Table 1: Selected variables: the categories for the 7 variables selected: 6 explanatory variables and 1 response variable.

(1)	Age at first diagnosis of MGUS	*AGE*
(2)	Sex of the patient	*SEX: 1 = male, 2 = female*
(3)	Type of plasma cell proliferative disorder	
	systemic amyloidosis	*AM*
	malignant lymphoproliferative disease	*LP*
	macroglobulinemia	*MA*
	multiple myeloma	*MM*
	no plasma cell proliferative disorder	*NO*
(4)	Albumin Level at MGUS diagnosis	*AL*
(5)	Serum Creatinine Level at MGUS diagnosis	*SCL*
(6)	Hemoglobin Level at MGUS diagnosis	*HL*
(7)	Size of Monoclonal Protein Peak at MGUS diagnosis	*SIZE*

The remainder of the paper is organized as follows. In section 2, we give principal tools of correspondence analysis (CA), including regression with MCA and a barycentric coding in MCA. Section 3 describes an overview on copulas based models. In section 4, we apply CA, binary logistic regression and copulas models on the MGUS data.

2 Exploratory Correspondence Analysis

2.1 Principal Tools of Correspondence Analysis

We will provide a concise summary of CA here, emphasizing geometrical and quantification aspects. For more geometrical details and proofs we refer to Benzécri [1], Cazes [4], Greenacre ([16], [17]), Lebart *et al*. [24], Le Roux *et al*. [25], Murtagh [27] and van der Heijden *et al*. [35].

CA is a technique with which it is possible to construct a multi-dimensional representation of the dependence between the row and column variables of a two-way contingency table. This representation is found by allocating scores to the row and column categories and displaying the categories as points, where the scores are used as coordinates (also called factors) of these points. These scores can be normalized in such a way that distances between row points and between column points in Euclidean space are equal to chi-square (χ^2) distances.

CA offers the remarkable feature of jointly representing individuals and variables. As a result of such analysis, not only does one gain insight in the relationship amongst individuals and amongst variables, but one can also find an indication of which variables are important in the description of which individuals. See Gordon [15].

Let *I* rows and *J* columns be collected into the $I \times J$ matrix *N* with elements n_{ij}. (For convenience of notation we will assume $I \geq J$ in this general discussion). A correspondence between these two finite sets *I* and *J* is defined by a function with positive integer values $n(i,j)$ on the product $I \times J$ what means to define a rectangular table $N_{IJ} = \{n(i,j): i \in I, j \in J\}$ of dimensions *card* (I) × *card (J)*.

Let $n_{i+} = \sum_{j \in J} n(i,j)$ and $n_{+j} = \sum_{i \in I} n(i,j)$ denote the sum of the *i*-th row and *j*-th column, respectively, and $n_{++} = \sum_{i \in I} \sum_{j \in J} n(i,j) = \sum_{j=1}^{J} \sum_{i=1}^{I} n_{ij} = 1^T N 1$ denote the grand total of *N*. The mass of the *i*-th row is $r_i = n_{i+}/n_{++} = p_{i+} = \sum_{j \in J} p_{ij}$ i.e. $r = P1$ and likewise the mass of the *j*-th column is defined as $c_j = n_{+j}/n_{++} = p_{+j} = \sum_{i \in I} p_{ij}$ i.e. $c = P^T 1$ where *P* is the so-called correspondence matrix of relative frequencies in proportion form defined as $P = N/n_{++}$ with entries $p_{ij} = n_{ij}/n_{++}$. In other words, the *i*-th row profile has mass equal to $p_{i+}$of that row in grand total. The mass center of the cloud of all row profiles is the centroid of the cloud and it is a profile that corresponds to the marginal row of *P*.

We proceed by considering chi-square distances between rows. These distances are computed on the profiles of the rows of a matrix, where the profile of row *i* is the vector of conditional proportions p_{ij}/p_i. The *"distributional distance"* also called χ^2 - distance between two rows *i* and i' between each profile and the centroid is measured by the chi-squared distance (χ^2) defined by

$$\delta^2(i,i') = \sum_j (p_{ij}/p_{i+} - p_{i'j}/p_{i'+})^2/p_{+j} \quad \text{where } p_{+j} \neq 0, \forall j \in J. \tag{2.0}$$

All of CA is based on *P* and the matrix *S* with elements $s_{ij} = (p_{ij} - r_i c_j)/\sqrt{r_i c_j}$. We note here that this makes CA invariant to rescaling of the original matrix *N*. In other words, the CA solution can be found as follows: Let *P* be the matrix to be analyzed. Let us construct $D_r = diag(r)$ and $D_c = diag(c)$, the diagonal matrices whose diagonal entries are respectively the marginal row proportions $p_{i+}$and column proportions p_{+j}, where it is assumed that $p_{i+} > 0$ and $p_{+j} > 0$.

Let *E* be the matrix with expected frequencies computed under the independence model. We can write $E = D_r t t^T D_c$, where *t* is a unit vector, the length of which depends on the context. Elements of *E* have the form

$$e_{ij} = p_{i+} p_{+j} = r_i c_j \tag{2.1}$$

The aim of the computational algorithm to obtain *factor coordinates* (or simply factors) of the row and column profiles with respect to principal axes, using the *singular value decomposition* (SVD), is as follows:

The matrix $S = D_r^{-\frac{1}{2}}(P - E)D_c^{-\frac{1}{2}}$ is submitted to SVD whose elements have the values $(p_{ij} - e_{ij})/\sqrt{e_{ij}}$; they are proportional to standardized residuals. These residuals are decomposed as follows:

$$D_r^{-\frac{1}{2}}(P - E)D_c^{-\frac{1}{2}} = U \Lambda V^T \tag{2.2}$$

where $U^T U = I = V^T V, \Lambda$ is the diagonal matrix of (positive) singular values λ_α in descending order: $\lambda_1 \geq \lambda_2 \geq \cdots, \alpha$ is the index for dimension and the matrices *U* and *V* are respectively the

row and column eigenvectors which consist of (left and right) singular vectors. Here, we define $D \leq \min(I-1, J-1)$, U is of order $I \times D$, V is of order $J \times D$ and Λ is of order $D \times D$ and is non-singular. These singular vectors are then used to compute respectively unweighted row and column scores called standard coordinates and normalized as follows:

$$R = D_r^{-\frac{1}{2}} U \tag{2.3a}$$

$$C = D_c^{-\frac{1}{2}} V \tag{2.3b}$$

Without going in the details, by substituting equations (2.3a) and (2.3b) into equation (2.2), we find

$$P = E + D_r R \Lambda C^T D_C = D_r (tt^T + R \Lambda C^T) D_C \tag{2.4}$$

which is known as a *"reconstitution formula"*. This formula shows that CA decomposes the departure from independence in matrix *P*.Whether this is the case or not can be tested by using the *Pearson's chi-squared statistic* of the data matrix, (i.e. the sum of squares of the matrix *S*) as follows:

$$\chi^2 = n_{++} \sum_{i \in I} \sum_{j \in J} (p_{ij} - e_{ij})^2 / (e_{ij}) \tag{2.5}$$

$$= n_{++} tr(SS^T) = n_{++} tr(S^T S) \tag{2.6}$$

where n_{++} is the sample size. The relation between χ^2 and the squared singular values in Λ^2 follows from equations (2.2) and (2.6) :

$$tr(\Lambda^2) = \chi^2 / n_{++} \tag{2.7}$$

In French publications $tr(\Lambda^2)$ is often called the "*total inertia*", a term from mechanics, which has however a precise meaning in CA, but does not have a sufficiently evocative power:

$$\Phi^2 = tr(SS^T) = tr(S^T S) = \sum_{i=1}^{I} \sum_{j=1}^{J} (p_{ij} - e_{ij})^2 / (e_{ij}) \tag{2.8}$$

In this context, the inertia is also the sum of squares of the singular values, i.e. the sum of the eigenvalues:

$$Inertia = \sum_{k=1}^{K} \lambda_k \tag{2.9}$$

where $K = \min\{I-1, J-1\}$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0$ are the nonzero eigenvalues of SS^T and *k* its rank, i.e. *k* = 1, 2, . . . ,*K*. Squares of singular values of *S* also decompose total inertia: $\lambda_1 \ldots \lambda_k$ are principal inertias.

The statistic χ^2/n is also often referred to as *"Pearson's index of mean-square contingency"*. Equation (2.8) shows that CA decomposes the chi-square value for testing independence in the matrix. χ^2 is *Pearson's chi-squared statistic* of the data matrix, i.e. the sum of squares of the matrix *S*.

When the input data is in *complete disjunctive form*, CA is termed MCA. Complete disjunctive form is a form of coding where the response categories, or modalities, of an attribute have one and only one non-zero response. MCA is defined as an extension of CA to more than $Q = 2$ variables, which allows one to analyze the pattern of relationships of several categorical dependent variables.

Suppose the original matrix of categorical data is $N \times Q$, i.e., N cases and Q variables. The first form of MCA converts the cases-by-variables data to an indicator matrix Z where the categorical data have been recorded as dummy variables. If the q-th variable has J_q categories, this indicator matrix will have $J = \sum_q J_q$ columns. Then the indicator version of MCA is the application of the basic CA algorithm defined above in sub-section 2. 1 to the matrix S, resulting in coordinates for the N cases and the J categories.

The second form of MCA calculates the $J \times J$ table obtained as $B = Z^T \times Z$ of all two-way cross-tabulations of the Q variables and is called the *"Burt table"*. Then the Burt version of MCA is the application of the same basic CA algorithm to the symmetric matrix B, resulting in coordinates for the J categories.

The standard coordinates of the categories are identical in both versions of MCA, and the principal inertias in the Burt version are the squares of those in the indicator version. Moreover, the eigenvalues obtained from CA of the Burt table give, in general, a better approximation of the inertia, explained by the factors, than the eigenvalues of Z.

2.2 Regression with MCA

We will provide a concise summary of Regression in the context of CA as proposed by Cazes ([6], [7], [8]) whose methodology is at the heart of MCA, studying the regression problem between a *response variable* and a set of *explanatory variables.* For more geometrical details and proofs we refer particularly to Cazes [4]. See also de Tibeiro and d'Ambra [10] and de Tibeiro [11].

We suppose that all the variables $x_1, x_2, \ldots, x_i, \ldots, x_p, y_1, y_2, \ldots, y_j, \ldots y_q$, have been divided into classes, and we designate by $K_{x_i}(resp. K_{y_j})$ the set of categories of the $x_i(1 \leq i \leq \mathrm{p})[\mathrm{resp}.y_j(1 \leq j \leq \mathrm{q})]$ and by $K_X(resp. K_Y)$ the unconnected union of $K_{x_i}(resp. K_{y_j})$, i.e. the set of all the explanatory categories (resp. to explain): $K_X = \cup\{K_{x_i} | i = 1, \ldots, p\}$; $K_Y = \cup\left\{K_{y_j} \middle| j = 1, \ldots, q\right\}$. If E designates the set of n observations, then we consider the complete disjunctive table (or indicator matrix) S_{EK_X} that we note simply S, associated with variables x_i of which the general term $S(e, k)$ is defined by:

$$S(e,k) = \begin{pmatrix} 1, \text{if } e \text{ has adopted the modality } k \text{ of} x_i, \\ 0, \text{if not.} \end{pmatrix} \qquad \text{for, } \forall e \in E, \forall k \in K_{x_i} \subset K_X,$$

We have the same notation for T_{EK_y}, (or simply T), the complete disjunctive table associated with variables y_j. We designate $T(e,k)(e \in E, k \in K_Y)$, as the general term of T.

Regression analysis with CA involves carrying out the following steps:

Step (1): After dividing into slices of variables x_i and y_j, we construct the table $C = T^tS$ (associated with the complete disjunctive table $t_{E(YJ)}$), which gathers together the set of q p contingency tables crossing every variable $x_i(1 \leq i \leq p)$ with every variable $y_j(1 \leq j \leq q)$.

Step (2): We carry out CA of data table *C*. We designate by $(\varphi_\alpha^{K_X}, \varphi_\alpha^{K_y})$ the α^{th} couple of factors associated with variance 1 (derived from this analysis) and by λ_α the corresponding eigenvalue.

Step (3): We add the table *S* to supplement *C*, *i*. *e*. we project on the *r* first factorial axes found in step (2) the profiles of the rows *e* in the table *S*. Let $F_\alpha(e)$ be the factors of the row's profile $e \in E$ on the factorial axis α.

Taking into account that $\sum\{S(e,k)|k \in K_X\} = p$, we obtain

$$F_\alpha(e) = \frac{1}{p}\sum_{k\in K_X} \varphi_\alpha^k S(e,k) \tag{2.10}$$

Moreover, it is well known that if one carries out CA of the complete disjunctive table *S* by keeping all the factors, the result of the regression is identical to the result of the analysis of variance. For more details, see Cazes [5] and Su-Myat [33]. See also de Tibeiro and d'Ambra [10].

2.3 Barycentric coding in Multiple Correspondence Analysis

Considering that we lose some information with coding in (0, 1), we propose in this paper to extend these considerations to *barycentric coding:* on the basis of *p* quantitative variables $x_j(j \in J)$ measured on a set *I* of *n* individuals. Let us note x_{ij} the value of the variable x_j for the *i*. Let us indicate by *X* the table of the $x_{i\ j}$. One gives oneself *r pivots*, i.e *r* values $t_1, t_2, \ldots, t_r$ $(t_1 < t_2 < t_3 \ldots < t_r)$.

If $(k(i,j_1), k(i,j_2), \ldots k(i,j_r))$ indicates the coding of x_j for the individual *i*, one poses:

If $x_{ij} \leq t_1$: $k(i,j_1) = 1;\ k(i,j_s) = 0$ if $s \neq 1$

If $x_{ij} \geq t_r$: $k(i,j_r) = 1;\ k(i,j_s) = 0$ if $s \neq r$

If
$t_m \leq x_{ij} \leq t_{m+1}$: $k(i,j_m) = (t_{m+1} - x_{ij})/(t_{m+1} - t_m);\ k(i,j_{m+1}) = (x_{ij} - t_m)/(t_{m+1} - t_m);$

$k(i,j_s) = 0$ if $s \neq m$ or $m+1$.

With this type of coding, where both of the values $k(i,j_s)$ are not null, there are the relations, for x_{ij} pertaining to the interval $[t_1,t_r]$: $\sum_{s=1}^{r}\{k(i,j_s)\} = 1$ and $\sum_{s=1}^{r}\{k(i,j_s)t_s\} = x_{ij}$ the barycentre of the affected points t_s of the masses $k(i, j_s)$. For more details, see Cazes [7], Benzécri ([2], [3]) and Gallego [12].

We propose indeed to apply this type of coding and then to carry out the usual linear regression with explanatory variables, of the factors resulting from MCA. To describe overall the connections between the variables, we propose to carry out CA of the Burt matrix with the *"barycentric table"* additional table. See Cazes ([4], [5], [7]) and Ouadrani [29].

Let q_r the response variable and Q_p the set of explanatory variables. We note:

$Q = \{q_r\} \cup Q_p$, $K = \{q_r\}$, $K^* = Q_p = Q - \{q_r\}$, $L = J_{q_r} = J_r$: the set of categories of the response variable,

$L^* = \cup \{J_q : q \in Q_p\} = J_p$: the set of the categories of *predictive* (or explanatory) variables. We note also

$J = J_{q_r} \cup J_p$.

To describe the connections between q_r and Q_p, we propose to subject to CA, the data set table $C_{J_r} \times J_p$ crossing J_r and J_p and we associate in additional supplementary with $C_{J_r} \times J_p$, the "barycentric table" evoked above $k_{I \times J_p}$ (instead of the disjunctive table as proposed by Cazes [4]), associated with the explanatory variables. One thus obtains on the maps resulting from CA a representation of the connections between the response variable and the explanatory variables, as a "visualized regression".

This visualization is only obtained starting from the knowledge of the explanatory variables, i.e., starting from the "barycentric table" $k_{I \times J_p}$, associated with these variables. Let us note $F_\alpha^I = \{F_\alpha(i) : i \in I\}$, the set of the coordinates of projections of elements $i \in I$ on the factorial axis α resulting from the table $C_{J_r} \times J_p$ called *"regression table"* or *"connection table"*.

Let us suppose that q_r comes from cutting in classes of a quantitative variable *y*. Then to explain *y*, we can carry out a usual regression on the factors F_α^I associated with the preserved factorial axes. A formula of regression of the type is thus obtained:

$y(i) \approx \sum \{b_\alpha F_\alpha(i) : \alpha \in A'\}$,where A' represents the set of preserved factors.

Let us suppose that one wants to do the prevision *y* for a new observation *s* for whom only the explanatory variables are known. It is enough to add *s* in supplementary row to with $C_{J_r} \times J_p$ (row of *r* and *I* following the explanatory categories taken by *s*) to have the $F_\alpha(s)$ (factor of *s* on the axis α) and to apply the preceding formula.

Also let us note that if we indicate by $(\varphi_\alpha^{J_r}, \varphi_\alpha^{J_p})$ the α^{th} couple of factors associated with variance 1 resulting from $C_{J_r} \times J_p$, we obtain:

$$F_\alpha(i) = \sum \{k(i,j)\varphi_\alpha^j : j \in J_p\}/Card\ Q$$

and thus the regression formula can be written in the form :

$y(i) \approx \sum \{d_j\ k(i,j) : j \in J_p\}/Card\ Q$ where $d_j = \sum \{b_\alpha \varphi_\alpha^j : \alpha \in A'\}$. For more details, see

Cazes ([4], [7]).

As $k(i,j)$ is worth 1, if *i* has adopted the modality *j* and 0 if not, the regression formula takes the following simple form: $y(i) \approx \sum\{d_{q(i)} : q \in Q_p\}/Card\ Q_p$ where *q*(*i*) indicates, let us recall it, the modality of J_q taken by *i*.

It can be thus satisfactory to note that a factor is narrowly correlated with a variable which does not appear explicitly in the data set table where it is resulting. It will be necessary to remember that an estimate of the parameters of the linear model becomes constraining when the explanatory variables of J_p are highly correlated linearly. To circumvent this difficulty, it is generally advised to resort to a *regression on principal components.* If the variables are very dependent not linearly, it is extremely useful to proceed by a *regression on the factors* resulting from MCA allowing to capture as well as possible the non-linear interrelations.

Finally, we note that when there are a large number of predictors, we can obtain a model with too many parameters, and one unavoidably duplicates the error. To avoid such an over-parametrization, the *Partial Least Square* (PLS) *regression* may be introduced. As the PLS components depend on the connection between the response variables and the predictors, we cannot calculate the variances of the regression coefficients with a simple formula. For more details, see Tenenhaus [34] and Cazes [5]. See also de Tibeiro and d'Ambra [10].

3 Copulas

Copulas are alternative probability measures of stochastic dependence and powerful tool for simulating joint probability distributions. Copulas are functions that join or couple multivariate distribution functions to their one-dimensional marginal distribution functions. Advantages of using copulas in modeling dependence are: (i) allowance to model both linear and non-linear dependence, (ii) arbitrary choice of marginal distributions and (iii) capable of modeling extreme endpoints. Among several contributions on copula based models and applications, few are mentioned by Clayton [9], Genest, Ghoudi and Rivest [13], Genest and MacKay [14], Herath and Kumar [18], Kumar [21, 22, 23], Nelson [28], Schweizer and Sklar [32].

The problem of specifying a probability model for independent bivariate observations $(x_1,y_1), \ldots, (x_n,y_n)$ from a population with distribution function *H*(*x*, *y)* can be simplified by expressing *H*(*x*, *y*) in terms of its marginal probability distributions *F*(*x)* and *G(y)*, and its associated dependence function, i.e., copula *C* implicitly defined through the identity: $C\big(F(x),G(y)\big) = H(x,y)$.

A natural way of analyzing bivariate data thus consists of estimating the dependence function and the marginal distributions separately. This two-step approach to stochastic modeling is often convenient, because many tractable models are readily available for the marginal distributions. See Plackett [26] and Clayton [9]. A more general solution to the problem of choosing an appropriate parametric family of dependence functions is using the *Archimedean family of copulas.* The basic assumption is that the data can be suitably modeled by an *Archimedean copula,* which implies that on the unit square, the appropriate dependence function is of the form:

$$C(u,v) = \phi^{-1}\{\phi(u) + \phi(v)\}, \qquad 0 < u, \qquad v < 1$$

for some convex decreasing function ϕ defined on (0, 1] in such a way that $\phi(1) = 0$.

By convention, $\phi^{-1}(t)$ is taken equal to 0 whenever $t \geq \phi(0)$. These conditions are necessary and sufficient for $C(u, v)$ to be a distribution function and are equivalent to the requirement that $1 - \phi^{-1}(t)$ be a unimodal distribution function on $[0, \infty)$ with mode at 0. For more details, see Schweizer and Sklar [28], Theorem 5.4.8. There are several copulas belonging to Archimedean family of copulas which have simple closed form expressions. See Nelsen [24].

Some examples of commonly used Archimedean bivariate copulas:

(i) Clayton copula: Copula generating function $\phi(t) = (t^{-\theta} - 1)/\theta$; Copula

$$C(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}, \qquad \theta \in [-1, \infty) \setminus \{0\}.$$

(ii) Gumbel copula: Copula generating function $\phi(t) = (-\ln t)^{\theta}$; Copula

$$C(u, v) = \exp[-\{(-\ln u)^{\theta} + (-\ln v)^{\theta}\}^{1/\theta}], \qquad \theta \in [1, \infty).$$

(iii) Frank copula: Copula generating function $(t) = -\ln \frac{e^{-t\theta} - 1}{e^{-\theta} - 1}$; Copula

$$C(u, v) = -\frac{1}{\theta} \ln \left[1 + \frac{(e^{-u\theta} - 1)(e^{-v\theta} - 1)}{e^{-\theta} - 1} \right], \qquad \theta \in (-\infty, \infty) \setminus \{0\}.$$

There are different approaches considered to estimation of copulas. Genest *et al.* [13] estimate a parametric family by maximum likelihood. See Genest and MacKay [14] for motivation, elementary properties, and convergence results concerning sequences of Archimedean copulas.

4 Sequence of the analyses

4.1. "Visualized Regression" or CA of the "Regression Table" $\mathbf{C}_{5\times27}$ resulting from the "regression table" $C_{J_r} \times J_p$ as described in section 2.2.

According to steps (1) and (2) of section 2.2., we have created a table $\mathbf{C}_{192\times27}(\mathbf{C}_{5\times27} \cup \mathbf{C}_{187\times27})$ whose 5 first rows are principal elements and the 187 last rows are supplementary elements. This additional table $\mathbf{C}_{187\times27}$ is precisely a complete disjunctive form associated with the 6 explanatory variables (*AGE*, *SEX*, *AL*, *SCL*, *HL*, *SIZE*).

Table 2: Principal inertias (eigenvalues)

dimension	principal inertia	% of inertia	cum %	scree plot
1	0. 037461	51. 393	51. 393	**********
2	0. 014469	19. 850	71. 243	****
3	0. 010686	14. 661	85. 904	***
4	0. 010275	14. 096	100. 000	***
Total	0. 072891	100. 000	100. 000	

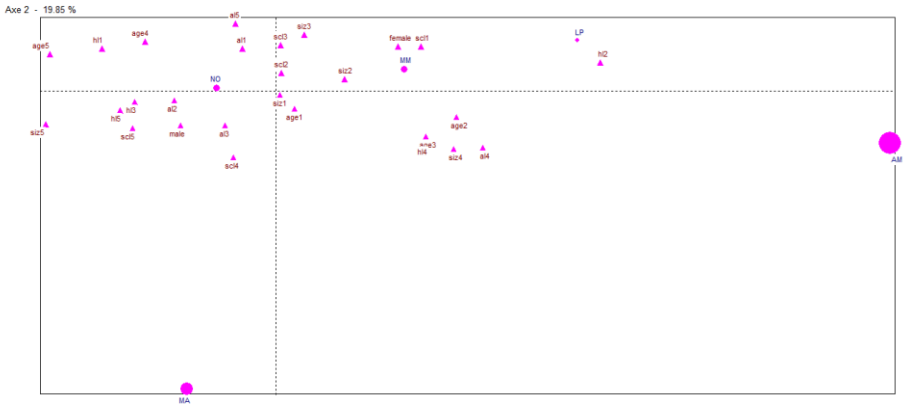

Figure 1: Correspondence Analysis factor map (1,2) of the "Regression Table" $C_{5\times27}$

- *Figure 1:* Cloud of 32 categories in the 1×2 plane. Five categorical response variables are represented in capital letters and 27 categorical explanatory variables are represented in small tiny letters.

The percentages of inertia explained by the top-ranked are 51.4%, 19.9%, 14.6% and 14.1%. Axis 1 therefore accounts for more than half of the total inertia of the cloud. Thus the first two dimensions account for almost 71.3% of the inertia.

The first factor F_1, clearly standing apart $(\tau_1 = 51.4\%;\ \tau_2 = 19.9\%)$, is a general attitude, a factor of general level i.e. the first eigenvalue which is associated to the first axis is more important than the three subsequent ones joined. We also note that the first three factors explain a little bit more than 86 % of the dispersion of observed values. We will admit that the essential part of structural links between the data is contained in the space of the first three dimensions.

According to factor projections, contributions and correlations, this first axis $(F_1 > 0)$ is the axis of the categorical response variable *AM* (systemic amyloidosis), a type of plasma cell proliferative disorder, which detaches itself from the other categorical explanatory and response variables. We observe secondarily on the same half plane $(F_1 > 0)$ four groups:

- A GROUP of categorical explanatory variables associated with the categorical response variable *MM* (multiple myeloma): a low Serum Creatinine level at MGUS diagnosis (*scl*1) and the gender *female* of the patient.

- A GROUP of categorical explanatory variables located in the fourth quadrant on the lower right: a high Albumin level at MGUS diagnosis (*al*4), a high level size of the Monoclonal Protein Peak at MGUS diagnosis (*siz*4), a high Hemoglobin level at MGUS diagnosis (*hl*4), and a low-median age at first diagnosis of MGUS (*age*2, *age*3).

- A GROUP of low Hemoglobin level at MGUS diagnosis (*hl*2) associated with the categorical response variable *LP* (malignant lymphoproliferative disease).

- A GROUP of categorical explanatory variables (*age*1, *siz*1, *siz*2, *scl*2, *scl*3, *siz*3) moving up, to the (0, 0) location at the centre of the display.

On the side of the negative first axis $(F_1 < 0)$, the categorical response variable *NO* (no plasma cell proliferative disorder) is associated with a high size of the Monoclonal Protein Peak at MGUS diagnosis (*siz*5), a high age at first diagnosis of MGUS (*age*5), a low hemoglobin level at MGUS diagnosis (*hl*1) and the gender *male* of the patient.

The interpretation of the first factor is then clear: *at MGUS diagnosis, more the age of the male patient is advanced* (*age*5: *more than* 74 *years), more the size of the monoclonal protein peak at MGUS diagnosis is raised* (*siz*5) *and less is the hemoglobin level* (*hl*1).

Axis 2 is dominated by the categorical response variable *MA* (type of plasma cell proliferative disorder called also "macroglobulinemia"), which detaches itself from the other categorical explanatory and response variables. The categorical response variable *MA* is opposed to all the categorical response variables except *AM*, whose contribution and correlation are negligible.

We have displayed the simultaneous representation (1, 2) plane of the categorical explanatory and response variables in *Figure 1* which is almost sufficient for the interpretation. It shows the samples spread out on a parabolic crescent which is known as the *Guttman effect.* The five points corresponding to the categorical response variables (*MA, NO, MM, LP, AM)* closely follow the parabolic curve. This is an index of a steep gradient within the data: these are arranged according to a series which is patently obvious not only on the axis 1 but in the plane 1×2. The first axis opposes the extreme values and the second one opposes the intermediate values to the extreme values. All the information is almost given by this first factor. We will admit here that the essential part of structural links between the data is contained in the space of the two first dimensions.

Let us follow the parabola of the samples from the negative extremity of the axis 1 to the other extremity. We find on the negative side $(F_2 < 0)$, the categorical explanatory variables *scl*4, *al*4, *male*, emanating from high level categories, and on the positive side $(F_2 > 0)$, the categorical explanatory variables *female*, *al*5, *siz*3.This means in other terms, that at MGUS diagnosis, *more the albumin level, and size of the monoclonal protein peak are advanced, more there are strong chances that relates to the female sex.*

According to step (3) in section 2.2., we perform an estimation of the response variable from the "Regression Table". One adds up the indicator matrix $S_{187\times27}$ [supplementary rows or vectors of description in (0, 1) of all the individuals of the basic sample] as supplementary to $C_{5\times27}$, while projecting on the first four (non-trivial) factorial axes found the profiles of the rows of table *S*. CA of the full Table $C_{192\times27}$obtained where each modality of the categorical response variable is regarded as a numeric variable, that one seeks to express in linear combination of the data variables, replaced here by the factors resulting from CA of the table in (0, 1) or of the "Regression Table". See Figure 2 where are projected the observations (patients) as supplementary points.

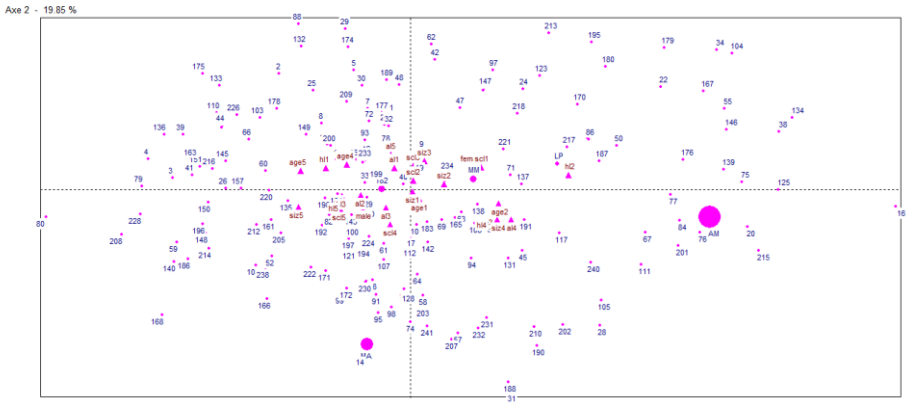

Figure 2: Correspondence Analysis factor map (1,2) of the table $C_{192\times27}$.

- *Figure 2:* Cloud of 219 categories in the 1×2 plane. Five categorical response variables are represented in capital letters, 27 categorical explanatory variables and 187 supplementary individuals are represented in small tiny letters.

These two maps (Figure 1 and Figure 2) clearly demonstrate a degree of separation between male $(F_1 < 0)$ and female patients $(F_1 > 0)$. There could exist some larger male-female differences. This, in fact, leads us in the following sub-section to carry out a separate study of the two groups: male and female. Let us announce however that in Figure 1, the lack of independence is evident in the fact that no categorical response variables except *NO* are plotted near the (0, 0) location at the centre of the display.

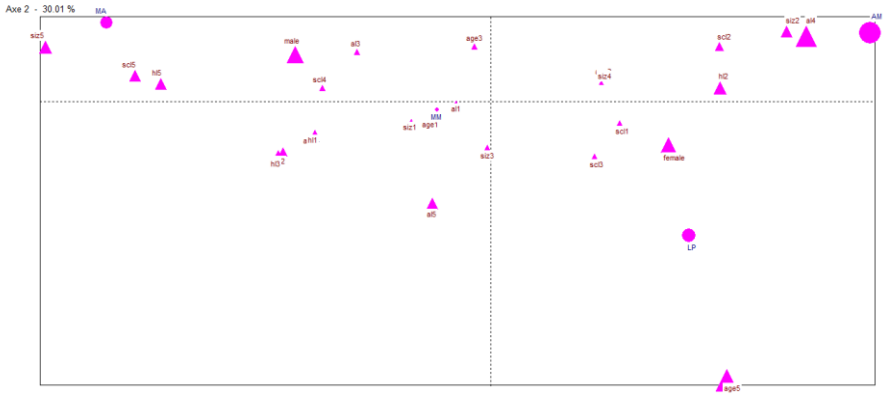

Figure 3: Correspondence Analysis factor map (1,2) of the table $C_{4\times27}$.

- *Figure 3:* Cloud of 31 categories in the 1×2 plane. Four categorical response variables are represented in capital letters and 27 categorical explanatory variables are represented in small tiny letters.

In Figure 3, we propose to superimpose *NO* to be predicted as a supplementary categorical response variable. This response category has no influence on the geometric orientation of the axes; rather, it supports and complements the interpretation of the configuration of active response categories. This point in Table 3, will have zero mass and thus plays no role in the analysis apart interpreting its position.

Table 3: Principal inertias (eigenvalues)

dimension	principal inertia	% of inertia	cum %	scree plot
1	0. 0936	43. 41	43. 41	**********
2	0. 0647	30. 01	73. 41	******
3	0. 0573	26. 59	100. 00	****

The percentages of inertia explained by the top-ranked are 43.41%, 30.01% and 26.59%. Axis 1 therefore accounts for more almost the half of the total inertia of the cloud. Thus the first three dimensions account for 100.00% of the total inertia.

The 1×2 plane which explains close to 75% of total inertia, opposes on the first axis the categorical response variables *AM* (*CTR* = 56. 9%) and *MA (CTR* = 29. 2%), located at the two extremities of the first two quadrants. On the side of the negative first axis $(F_1 < 0)$, we have noted the proximity of the categorical response variable *MA* with categorical explanatory variables: *siz*5 (high size of the Monoclonal Protein Peak at MGUS diagnosis), *scl*5, *scl*4 (high Serum Creatinine Level at MGUS diagnosis of MGUS), *hl*5 (high hemoglobin level at MGUS diagnosis and the gender *male* of the patient.

On the side of the positive first axis $(F_1 > 0)$, the categorical response variable *AM* is associated with a high Albumin Level at MGUS diagnosis (*al*4), a low size of the Monoclonal Protein Peak at MGUS diagnosis (*siz*2), a low Serum Creatinine Level at MGUS diagnosis (*scl*2), a low hemoglobin level at MGUS diagnosis (*hl*2) and the gender *female* of the patient.

These two categorical response variables (*AM* and *MA*) in opposition on the first axis are now opposed to the other response variables except the categorical response variable *MM* located at the centre of gravity of the cloud, are now associated on the half plane $(F_2 > 0)$ in opposition with the categorical response variable *LP*, associated with (*age*5: more than 74 years), a high age at first diagnosis of MGUS.

4.2 "Visualized Regression" or CA of the "Regression Tables" $rt_{109\times15}$ **and** $rt_{86\times15}$ resulting from the "regression table" $C_{J_r} \times J_p$ as described in section 2. 2.

We are interested in the relationship within a set of variables from the original table $\mathbf{n}_{187\times7}$ where 187 is the number of observations and 7 is the number of the variables retained in section 2 : One response variable *TYPE* (*AM*, *LP*, *MA*, *MM*, *NO*) and 6 explanatory variables (*AGE*, *SEX*, *AL*, *SCL*, *HL*, *SIZE*). See Table 1 in section 1.

The division of this data set into two groups: men and women, before an analysis is completed, reveals *a priori* that the categorical response variable *LP* (malignant lymphoproliferative disease) is not related to the gender "male". In a similar way, the categorical response variable *MA* (macroglobulinemia) is not related to the gender "female".

4.2.1 "Visualized Regression" or CA of the "Regression Table" $rt_{109\times15}$

From this table $\mathbf{n}_{187\times7}$, we consider first the table 105×6 crossing the 105 male patients and all the explanatory variables except *SEX* and the response variable *TYPE*. Alternatively, from the "responses" of the patients, we construct the "barycentric" table $Z_{105\times19}$ crossing the 105 male patients and the $4 + (5 \times 3)$ response categories. More precisely, we create the table

$Z_{105\times19}$ where the 4 first columns *AM*, *MA*, *MM*, *NO* are categorical response variables and the 15 following column categories are associated to the five explanatory variables (*AGE*, *AL*, *SCL*, *HL*, *SIZE*).

We believe more useful to subject the 6 variables retained (5 explanatory and 1 response variables) of the original data set to a "barycentric" coding according to the *"the personal equation"*. In each block J_q, an individual *i* has generally non null notes in two successive modalities. That means, *i* occupies an intermediate position. Thus, the coding of each notation *j* according to three categories {*j* +, *j* =, *j* -}, from which is built a generalized Burt matrix $B_{19\times19} = Z^t_{19\times105} \times Z_{105\times19}$. For more details, see Benzécri ([2], [3]) and McGibbon *et al.* [26].

From the contingency table $B_{19\times19}$, we take the sub-table $B_{4\times15}$ on which we add the table $Z_{105\times15} \subset Z_{105\times19}$. We obtain a "Regression Table" $rt_{109\times15}$. For more details, see Cazes [4] and Su-Myat [33].

General view of the results

The eigenvalues and the percentages of inertia (in parenthesis) associated to the four first axes are relatively small: $\lambda_1 = 0.046489$ $(\tau_1 = 49.098\%)$; $\lambda_2 = 0.039397$ $(\tau_2 = 41.608\%)$ and $\lambda_3 = 0.008800$ $(\tau_3 = 9.294\%)$.

According to factor projections, contributions and correlations, the first axis indicates a cleavage between Systemic Amyloidosis *AM* and Macroglobulinemia *MA* which contribute most to this axis (52.5% and 45.6%), respectively.

On the side of the negative first axis $(F_1 < 0)$, *MA* is particularly associated with the categorical explanatory variables: $SIZE+$ and $AGE-$. On the side of the positive first axis $(F_1 > 0)$, *AM* is particularly associated with $SCL-$. *MA* has higher association with $SIZE+$ and $AGE-$. This means in other terms that *the patient has the higher risk of possible development of serious plasma cell proliferative disorders when the size of the monoclonal protein peak at MGUS diagnosis is advanced and the age of the male patient is less.*

The second factor contrasts the categorical response variable *NO*, with the categorical response variables *AM* and *MA*. The categorical response variable *NO* is related particularly to the categorical explanatory variable $AGE+$. The categorical response variables *AM* and *MA*, located on the positive side of this axis, are not particularly associated with the other categorical explanatory variables.

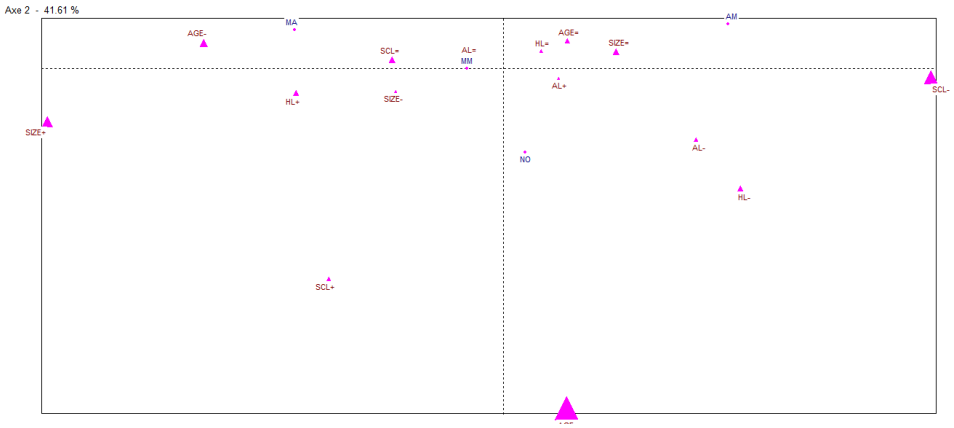

Figure 4: Correspondence Analysis factor map (1,2) of the "Regression Table" $rt_{109\times15}$.

- *Figure 4. H-Flou-TPMGUS*

Table 4: First Group: 105 *male patients. The categories for the* 5 *explanatory variables and* 1 *response variable.*

Variable	*Levels*	*Variable*	*Levels*
variable *AGE*	$j-, j =$ and $j+.$	variable *HL*	$j-, j =$ and $j+.$
variable *AL*	$j-, j =$ and $j+.$	variable *SIZE*	$j-, j =$ and $j+.$
variable *SCL*	$j-, j =$ and $j+.$	variable *TYPE*	$j-, j =$ and $j+.$

4.2.2 "Visualized Regression" or CA of the "Regression Table" $rt_{86\times15}$

From table $\mathrm{n}_{187\times7}$, we consider the table 82×6 crossing the 82 female patients and all the explanatory variables except obviously *SEX* and the type of plasma cell proliferative disorder. Alternatively, from the "responses" of the patients, we construct the "barycentric" table $Z_{82\times19}$ crossing the 82 female patients and the $4 + (5 \times 3)$ response categories. More precisely, we create the table $Z_{82\times19}$ where the 4 first columns *AM*, *LP*, *MM*, *NO* are categorical response variables and the 15 following columns categories are associated to the explanatory variables (*AGE*, *AL*, *SCL*, *HL*, *SIZE*). From this table $Z_{82\times19}$, we retain the sub-table $B_{4\times15}$ on which we add the table $Z_{82\times15}$ included in $Z_{82\times19}$. We obtain a "Regression Table" $rt_{86\times15}$.

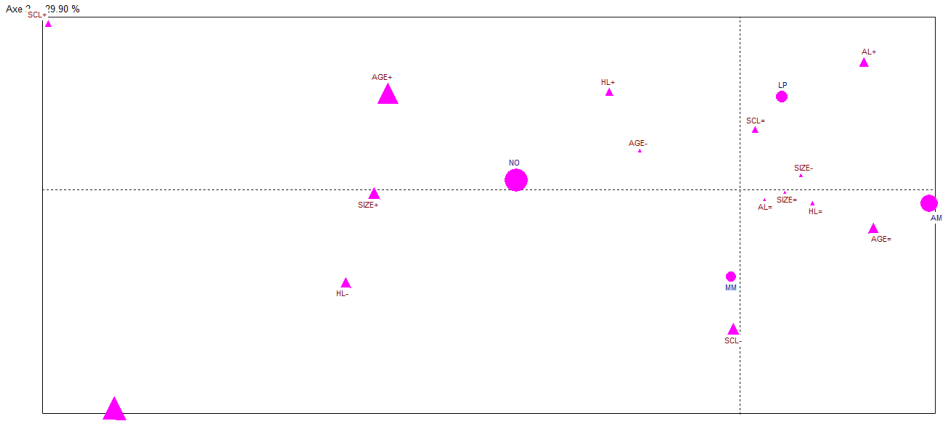

Figure 5: Correspondence Analysis factor map (1,2) of the "Regression Table" $rt_{86\times15}$.

- *Figure 5. F-Flou-TPMGUS*

Table 5: Principal inertias (eigenvalues)

$\lambda_1 = 0.030052$ $\tau_1 = 63.268\%$
$\lambda_2 = 0.014202$ $\tau_2 = 29.900\%$ $\tau_1 + \tau_2 = 93.168\%$
$\lambda_3 = 0.003245$ $\tau_3 = 6.832\%$ $\tau_1 + \tau_2 + \tau_3 = 100\%$

Thus, the axis 1 represents more than half of the total inertia of the cloud. This axis indicates a contrast between *NO* and *AM* (Systemic Amyloidosis). On the side of the positive first axis $(F_1 > 0)$, *AM* is particularly associated with *AGE* =. On the side of the negative first axis $(F_1 < 0)$, *NO* is related to the predictive categories: $AL -, AGE +, SIZE +$ and $HL -$.

Here we confirm again a result of CA of "Regression Table" $rt_{109\times15}$ in the sub-section 4.2.1: $AGE =$, the age between the two extremes is the only predictive category being able to have a relatively important relationship with *NO*, the absence of any of the clinical signs of malignant MGUS.

Table 6: Second Group: 82 female patients. The categories for the 5 explanatory variables and 1 response variable.

Variable	*Levels*	*Variable*	*Levels*
variable *AGE*	$j-, j =$ and $j+.$	variable *HL*	$j-, j =$ and $j+.$
variable *AL*	$j-, j =$ and $j+.$	variable *SIZE*	$j-, j =$ and $j+.$
variable *SCL*	$j-, j =$ and $j+.$	variable *TYPE*	j_a, j_b, j_c and $j_d.$

4.3 Logistic Regression Analysis

4.3.1 Binary Logistic Regression

We have run the preliminary analysis including estimating marginal distributions of the predictive variables (*AL*, *SCL*, *HL*, *SIZE*). Parametric and non-parametric, both, measures of correlation $(r = 0.426;\ \text{Kendall's}\ \tau = 0.272;\ \text{significant at 0.01 level})$ suggest that pair (*AL*, *HL*) is significant and rest of pairs are not.

Three parametric forms of distributions which fit close to these data are:

(i) $AL \sim N\ (3.206;\ 0.4739);\ HL \sim N\ (13.153;\ 1.7369)$

(ii) $AL \sim logN(3.169;\ 0.155);\ HL \sim logN(13.025;\ 0.145)$

(iii) $AL \sim \Gamma(45.773;\ 14.278);\ HL \sim \Gamma(57.347;\ 4.360)$

We can consider (*AL*, *HL*) as variables of interest. Using copulas, we will use Hemoglobin Level (*HL* at MGUS diagnosis) to be the predictor variable which makes a better sense in the context of this data set.

4.4 Copulas

4.4.1 Prediction of probabilities that patient has plasma cell proliferative disorder (Y) given *AL* levels (X)

We have estimated the prediction probabilities of patients with plasma cell proliferative disorder given *AL* level from the data set of 187 patients and the 50 Gumbel simulations of (*AL* and *Y* = 1, if plasma disorder present; else 0, if plasma disorder absent).

Summary statistics of the given data set: *n* = 187, mean (*AL*) = 3.2, mean (*Y*) = 0.2353, standard deviation (*AL*) = 0.4739, standard deviation (*Y*) = 0.4253, correlation *r* (*AL*, *Y*) = 0. 089, Kendall's $\tau(AL, Y)$ = 0.06. Estimated Marginal distributions: *Y* ~ Bernoulli (ρ = 0.2352) and *AL* ~ Gamma (45.773, 0.07). Estimated copula parameters: θ (Gumbel) = 1. 0650, θ (Clayton) = 0.1299, θ (Frank) = 0. 5518.

Distance from empirical copula: (Gumbel) = 1.581, Clayton = 1.636, Frank = 1.600. Which is the most appropriate copula in this case? Gumbel copula since minimum distance = 1.581.

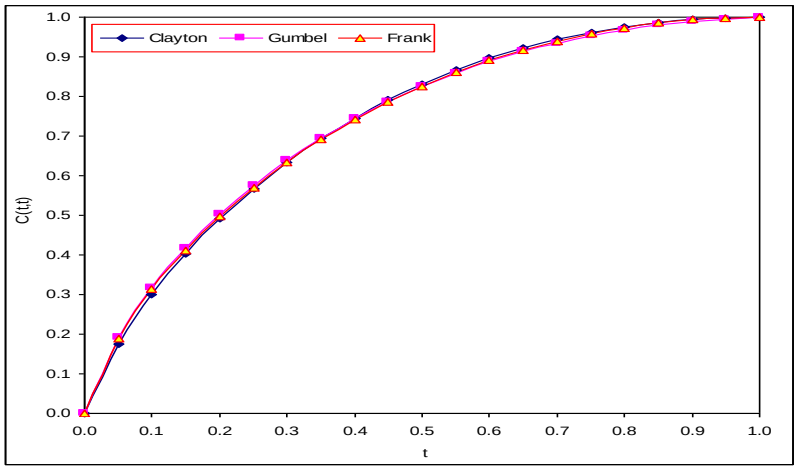

Figure 6: Clayton, Gumbel and Frank coplas.

- *Figure 6: Clayton, Gumbel and Frank copulas*

Plasma cell disorder prediction model estimated from data *n* = 187

$$\log it[P(Y = 1)] = -2.636 + 0.451 \times AL$$

Predicted probability that a patient will have plasma cell disorder

$$P[Y = 1] = [1 + \exp\{-(-2.636 + 0.451 \times AL)\}]$$

Thus, the probability that a patient who have *AL* level at 3 will have plasma cell proliferative disorder = 0.2170. Plasma cell disorder prediction model estimated from Gumbel copula

$$\log it[P(Y = 1)] = -2.746 + 0.4643 \times AL$$

and the predicted probability that a patient will have plasma cell disorder

$$P[Y = 1] = [1 + \exp\{-(-2.746 + 0.4643 \times AL)\}]$$

Figure 6 presents the predicted probabilities that patient will have plasma cell proliferative disorder given *AL* levels. These probabilities are based on fifty Gumbel copula simulations.

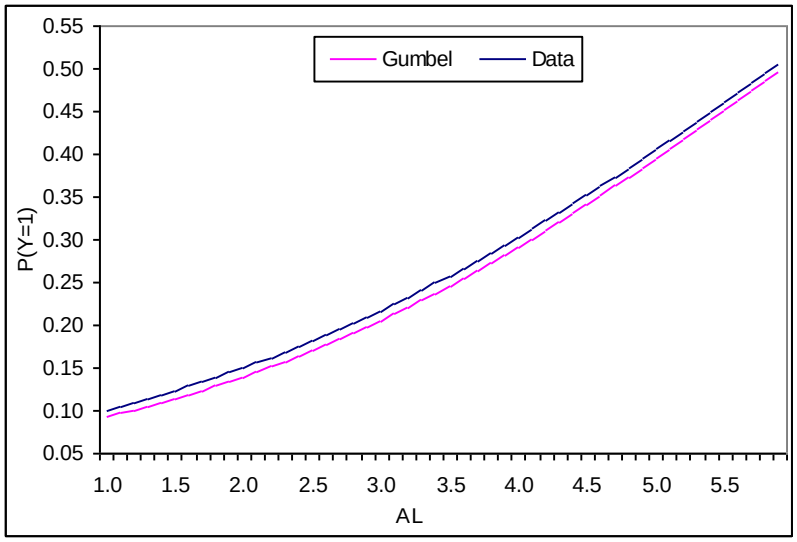

Figure 7: Predicted probabilities based on Gumbel copula.

- *Figure 7: Predicted probabilities based on Gumbel Copula.*

4.4.2 Link between the copula results and previous CA's results

We have already discussed results from MCA. Now referring to how to connect the copula results in the context of present data analysis, we have indicated from MCA results:

(i) The more the age of male patient is advanced; the more the size of Monoclonal Peak at MGUS is raised. This confirms the results found in the Binary Logistic Regression.

(ii) *MA* is associated with explanatory variables gender, *AL* and *SIZE*. One way of introducing copula in this context could be to consider prediction of size at MGUS using age as the explanatory variable for gender male. Thus it may show a relevant connection.

An important issue in prediction modeling of multivariate data is the measure of dependence structure. The use of Pearson's correlation as a dependence measure has several pitfalls and hence application of correlation models may not be an appropriate methodology. As an alternative, a copula based methodology for prediction modeling and an algorithm to simulate data are useful. This algorithm based on the marginal distributions of random variables is applied to construct the *Archimedean copulas. Monte Carlo* simulations are carried out to replicate data sets, estimate prediction model parameters.

We will continue later the validation of the prediction model by *Lin's concordance measure.* From skewness and kurtosis values, there is an indication that both age and size variables are

slightly skewed negatively and positively respectively and hence have some departures from symmetry.

5 Concluding Remarks: Perspectives, Limitations and Interest of the Study

The main thrust of this research project is to be found in the duality, the "cohabitation" and the complementarity of the exploratory and modeling approaches including some models based on Archimedean copulas as the appropriate measure of association.

More than just a simple pleasure of discovering results, at the end of a simulation computation, we expected to connect a "functional model of continuous correspondence" through the "regression table" with the binary logistic regression and barycentric linear coding.

Traditionally, MCA has been used prevalently on categorical data in the social sciences, but its application has been extended also to (positive) physical quantities. We have shown that MCA applied to medical data provides as informative and concise means of visualizing this data, a capacity for revealing relationships both among either patients or laboratory continuous values (variables) and between patients and variables.

Visualization by using MCA is based on representing χ^2 distance among "individuals" and variables, thus representing a decomposition of the value of the χ^2 statistic. Emphasis is placed on the "individuals" and variables that contribute to this value through their association.

In this respect, the use of a "Regression Table" for MCA to analyze the type of plasma cell proliferative disorder for MGUS revealed an excellent discrimination according to the sex and the age of the patients accidentally discovered during the process of being examined for other indications. *More precisely, the greater the male patient's age (more than 74 years), the larger the size of monoclonal protein peak at MGUS diagnosis and less the Hemoglobin level.*

According to the *p-values* obtained from the model $Status = f(age, AL, SCL, HL, size)$ of Binary Logistic Regression (containing all explanatory variables except sex), we find that age and size are the most interesting variables.

From this result, we could propose, as an alternative approach, some models based on the currently popular idea of Archimedean copulas as an appropriate measure of association. For illustration, we introduced copulas in this context and estimated prediction model for predicting the size of MGUS using age as the explanatory variable for the male gender. However copulas are applicable to the multivariate data situations as well which will be considered somewhere else in future.

Acknowledgment

This work is partially supported by New-Brunswick Innovation Foundation (NBIF). The second author is grateful to Pr. Bruce Jones, chairman of the Department of Statistical & Actuarial Sciences of the University of Western Ontario (UWO, Canada). Pr. Jones and all my colleagues in this department have always encouraged me to continue our professional relationship by maintaining my standing as Adjunct Research Professor, by providing continuous support of my research projects with their collegiality, through the courses I have taught and the Masters students that I have been lucky enough to guide. He wishes to thank also Pr. Pierre Cazes

(Université de Paris-Dauphine, France) for his helpful comments on Regression Analysis centered on Barycentric coding. Special thanks to Pr. Duncan Murdoch (UWO, Canada) for his help in finding the dataset for this study.

References

[1] J. -P. Benzécri, *Correspondence Analysis Handbook*. Marcel Dekker, (1992).

[2] J. -P. and F. Benzécri, Le codage linéaire par morceaux : réalisations et applications. *Les Cahiers de l'Analyse des Données*, 14 (2) (1989a), 203-210.

[3] J. -P. and F. Benzécri, Codage linéaire par morceaux et équation personnelle. *Les Cahiers de l'Analyse des Données*, 14 (3) (1989b), 331-336.

[4] P. Cazes, Analyses des données approfondies : Notes de cours du département de mathématiques et informatique de la décision et des organisations. *Université Paris Dauphine*, (2006-2007).

[5] P. Cazes, Adaptation de la régression PLS au cas de la régression après l'analyse des correspondances multiples. *Revue de Statistique Appliquée*, 45 (21), (1997), 89-99.

[6] P. Cazes, Méthodes de régression : Polycopié de 3ème cycle. *Université Paris Dauphine*, (1996).

[7] P. Cazes, Codage d'une variable continue en vue de l'analyse des correspondances. *Revue de Statistique Appliquée*, 38 (3) (1990), 33-51.

[8] P. Cazes, L'École d'été du CNRS sur l'analyse des données : Régression. *Laboratoire du Pr. J.-P. Benzécri, Université Pierre et Marie Curie* (Paris VI), (1977).

[9] D. G. Clayton, A Model for Association in Bivariate life tables and its application in Epidemiological studies of familial tendency in Chronic disease incidence. *Biometrika*, 65 (1) (1978), 141-151.

[10] J. J. S. de Tibeiro, and L. d'Ambra, An integrated approach to Regression Analysis using Correspondence Analysis and Cluster Analysis. *Statistica & Applicazioni*, 8 (1) (2010), 1-32.

[11] J. J. S. de Tibeiro, Consommation d'électricité sous un climat extrême : Estimation en fonction de la date et de la température. *Les Cahiers de l'Analyse des Données*, 22 (2) (1997), 199-210.

[12] F. J. Gallego, Codage flou en analyse des correspondances. *Les Cahiers de l'Analyse des Données*, 7 (4) (1982), 413-430.

[13] C. Genest, K. Ghoudi and L. -P. Rivest, A semiparametric estimation procedure of dependence parameters in multivariate families of distribution. *Biometrika*, 82 (3) (1995), 543-552.

[14] C. Genest and R. J. MacKay, Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *Canadian Journal of Statistics*, 14 (2) (1986), 145-159.

[15] A. D. Gordon, *Classification*. Chapman and Hall, 2nd edition, (1999).

[16] M. J. Greenacre, *Correspondence Analysis in Practice*. Second Edition, Chapman & Hall/CRC, (2007).

[17] M. J. Greenacre, *Theory and Applications of Correspondence Analysis*. Academic Press, (1984).

[18] H.S.B., Herath and P. Kumar, Research directions in engineering economics-modeling dependencies with copulas , *Engineering Economist*, (45) (1) (2007), 1-36.

[19] International Myeloma Working Group, Criteria for the classification of monoclonal gammopathies, multiple myeloma and related disorders: a report of the International Myeloma Working Group. *Br. J. Haematol*, 121 (5) (2003), 749–57.

[20] R. A. Kyle, "Benign" monoclonal gammopathy - after 20 to 35 years of follow-up. *Mayo Clinic Proceedings*, (1993), 6826-6836.

[21] P. Kumar, Statistical Dependence: Copula functions and mutual information based measures, *Journal of Statistics Applications & Probability: An International Journal*, 1(1) (2012), 1-14.

[22] P. Kumar , Copulas: Distribution functions and simulation, In Lovric, Miodrag (Ed*), International Encyclopedia of Statistical Science*, Springer Science +Business Media, LLC, Heidelberg (2011).

[23] P. Kumar , Probability distributions and estimation of Ali-Mikhail-Haq Copula, *Applied Mathematical Sciences: Journal for Theory & Applications*, (4) (2010), 657-666.

[24] L. Lebart, A. Morineau and K. M. Warwick, *Multivariate Descriptive Statistical Analysis, Correspondence Analysis and Related Techniques for Large Matrices*. John Wiley & Sons, Inc., (1984).

[25] B. Le Roux and H. Rouanet, Geometric Data Analysis: *From Correspondence Analysis to Structured Data Analysis*. Dordrecht, Kluwer, (2004).

[26] B. McGibbon Taylor, P. Leduc and J. J. S. de Tibeiro, Analyse des réponses des étudiants à un questionnaire relatif au mémoire de recherche de la maîtrise en administration des affaires. *Les Cahiers de l'Analyse des Données*, 14 (3) (1989), 337-346.

[27] F. Murtagh, *Correspondence Analysis and Data Coding with Java and R*. Chapman & Hall/CRC, (2005).

[28] R. B. Nelsen, *An Introduction to Copulas: Lecture Notes in Statistics*. Springer, New York, (2006).

[29] El. A. Ouadrani, Généralisation du tableau de Burt et de l'analyse de ses sous-tableaux dans le cas d'un codage barycentrique. *Les Cahiers de l'Analyse des Données*, 19 (2) (1994), 229-246.

[30] R. L. Plackett, A class of bivariate distributions. *Journal of American Statistical Association*, 60 (2) (1965), 516-522.

[31] [31] R. Rousseau, B. Augereau, A. Daver and D. Leguay, Méthodologie de la régression et de la prédiction fondée sur la classification automatique. *Les Cahiers de l'Analyse des données*, 16 (4) (1991), 479-488.

[32] B. Schweizer and A. Sklar, *Probabilistic Metric Spaces*. Elsevier, North-Holland, New York, (1983).

[33] K. K. Su-Myat, Multivariate Analysis Approaches: An application for Monoclonal Gammopathy of Undetermined Significance (MGUS). Master's thesis, *Dept. of Statistical & Actuarial Sciences, The University of Western Ontario*, (2008), 1-59.

[34] M. Tenenhaus, *La Régression PLS, Théorie et Pratique*. Technip, Paris, (1998).

[35] P. G. M. van der Heijden, A. de Falguerolles and J. de Leeuw, A Combined Approach to Contingency Table Analysis using Correspondence Analysis and Log-linear Analysis. *Applied Statistics*, 33 (2) (1989), 249-292.