

**USING THE CENTROID FOR VARIABLE SELECTION WITH MISSING DATA**

**Author's name(s):** Gaudart J*¹, Giustiniani J², Leonetti G².

**Affiliation(s):** ¹Laboratoire de Biomathématiques et Statistiques Médicale, Faculté de Médecine, Université Aix-Marseille II, 27 bd J. Moulin 13005 Marseille, France.

²Laboratoire de Médecine Légale, Faculté de Médecine, Université Aix-Marseille II, 27 bd J. Moulin 13005 Marseille, France.

**E-mail:** jean.gaudart@medecine.univ-mrs.fr

**Phone:** +33(0)491791910 and +33(0)491324516; **Fax:** +33(0)491794013 and +33(0)491324512

**Corresponding author:** Jean Gaudart

**Keywords:** Data analysis, variable selection, missing data.

**Topic area of the submission:** Model selection, Data mining

**Abstract**

**Introduction:** In life science, all variables are not always fully observed, since subject recruitment and data measurements are difficult. Particularly in anthropology and legal medicine, the number of variables is often greater than the recruited sample, because of missing data. Then, statistical analysis and result interpretation are tricky. Indeed, stepwise regression methods, Partial least Square regression methods [1] or classical data analysis methods are not able to select variables. Then, in this context, variables are often selected with arbitrary or subjective means.

For a homogeneous set of variables, we propose selecting the nearest variable to the centroid of the variable set. This variable minimizes the distance to the other variables, and therefore an important part of the set information is taken into account.

The aim of this study was to develop a non arbitrary method based on the centroid of a homogeneous variable set.

**Material and method:** In case of abortions caused by traumatic situations, the fetal age has to be accurately determined. Indeed, in the French law, the judgment and the prejudice appreciation differ in step with the fetal age in those situations. Fetal age can be determined using thigh-bone lengths [2] or foot lengths [3], as well as cranium bones measures. For the
latter determination, 46 biometric variables were measured (cranium bone lengths or angular measures [4, 5]) on 50 aborted fetuses. But, because of the traumatic situations, it was not possible to measure all the variables on the whole fetus sample. The measures were complete for only 18 fetuses.

Firstly, to obtain homogeneous variable sets, we used a hierarchical cluster analysis using Pearson correlations as similarity measures. Secondly, for each obtained variable set, we selected the nearest variable to the centroid of the set. We have determined that the nearest variable to the centroid is the variable which maximizes the sum of correlations. Indeed, the squared distance between a variable and the centroid is given by the Torgerson formula [6, 7]:

$$\delta^2(y, j) = \frac{\delta^2(j,..)}{p} - \frac{1}{2} \frac{\delta^2(\cdot, \cdot)}{p^2},$$

where $$\delta^2(j,..) = \sum_k \delta^2(j, k)$$ and $$\delta^2(\cdot, \cdot) = \sum_j \sum_k \delta^2(j, k)$$ and $$p$$ is the number of variables.

If $$\delta^2(j, k) = 2(1 - \rho_{jk})$$ is the squared distance between two variables, where $$\rho_{jk}$$ is the correlation coefficient between two variables, then $$\min_j \delta^2(y, j) \equiv \max_j \sum_k \rho_{jk}$$

Thirdly, with the variables selected in this way, the sample size was sufficient to compute a stepwise linear regression model. The results were analyzed by cross-validation procedure.

In the presented statistical context (finite sample size and $$\rho_{jk} \neq 0$$), the distribution of the empirical correlation coefficient is not simply usable [8]. Then, it is not possible to formalize the distribution of the maximum of the empirical correlation sum ($$C_g$$). Therefore we studied the empirical distribution of $$C_g$$ by simulation.

**Results:** The hierarchical cluster analysis, using 46 variables, built 2 sets with respectively 36 and 4 clustered variables. The 6 remaining variables appeared to be independent of the other variables. In the first set (36 variables) we selected the variable called mLGB using the $$C_g$$ criterion. That is, this variable maximized the correlation sum and therefore was the nearest variable to the centroid of the first set. The variable mLGB represents the length between the frontal bone and the foramen magnum. This length is a classical and standardized measure [9] and, for fetuses, mLGB can be easily measured by scanner. In the same way, we selected the variable called mAPA (alveolar profile angle [5, 9]) in the second set (4 variables).
On the whole, we obtained 8 variables, with a sample size of 35 fetuses (free of missing data), to compute a stepwise linear regression model. We obtained the following model:

\[ age = -4.591 + 0.692 \times mLGB \]

(age weeks of amenorrhea, mLGB in millimeter).

The model was adequate (\(R^2=0.9\)) and the cross-validation showed good predictions (mean and 95% confidence interval of the prediction errors: \(1.3 \times 10^{-3} [-0.582; 0.584]\)).

With 1000 simulations (figure 1), the empirical distribution of \(C_g\) was stable. The empirical mean (31.91 CI95% [31.86; 31.96]) and the empirical median (32.07) were close to the theoretical value (32.03) in spite of the small size of the simulated samples (\(n=19\) subjects). The empirical variance (0.56) and the empirical range (minimum=26.75; maximum=33.22) were also small.

**Figure 1**: empirical distribution of \(C_g = \max_j \sum_k r_{jk}\),

1000 simulations of 36 Normal variables.

**Discussion**: When the number of continuous variables is greater than the sample size, because of missing data process, all other procedure are subjective. In this context, the choice of the nearest variable to the centroid of a variable set is a non arbitrary and simple method. Moreover the almost sure convergence of the selection criterion is easily proven.
For detecting homogeneous variable sets we have chosen hierarchical cluster analysis. But, other clustering procedures can be used, such as K-means clustering. In our study, in spite of small changes in the variable sets, using K-means clustering did not change the final prediction model: the mLGB variable was also the only variable in the final model. But in the K-means method the number of cluster has to be fixed a priori. On the contrary, the number of clusters can be chosen a posteriori in the hierarchical clustering, according to the results of the similarity.

Missing data do not disturb our selection method, as far as the missing process is independent of the variables. Indeed, the correlation is estimated between two variables, and for these two variables the number of data has to be sufficient to ensure the convergence of the correlation estimation.

The maximum of the Pearson correlation sum is a useful criterion, and its practical use is very simple and fast. Thus, the choice of the nearest variable to the centroid is a non arbitrary selection method usable in the presence of missing data.

Acknowledgments: The authors gratefully acknowledge Bernard Fichet for his detailed and helpful comments.

References: